模型压缩

汇报人：扣扣

大纲

- 训练和部署之间的区别
- 参数量化（quantization）
- 模型裁剪/剪枝（Pruning）
- 模型结构重构
- 知识蒸馏

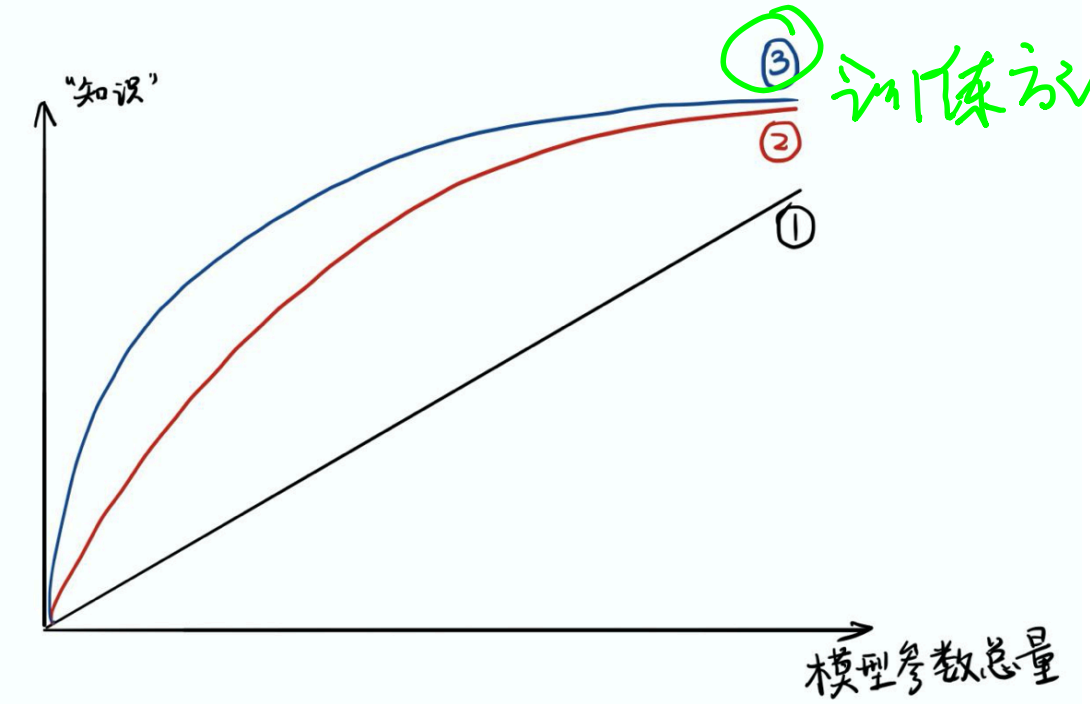
NeuHub
AI Links All

训练和部署之间的区别

大模型不方便部署到服务中去，常见问题如下:

- 推断速度慢
- 对部署资源要求高(内存，显存等)

训练和部署之间的区别



NeuHub
AI Links All

贪心科技 GREEDY TECHNOLOGY

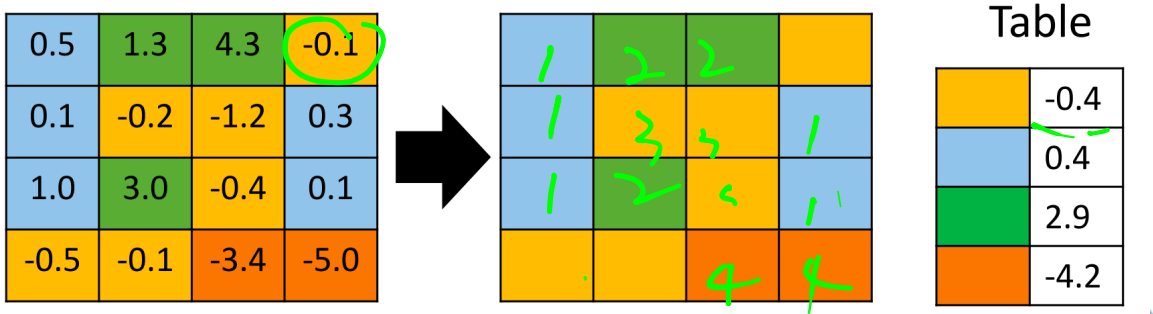
京东智联云 & 贪心科技 GREEDY TECHNOLOGY

参数量化（quantization）

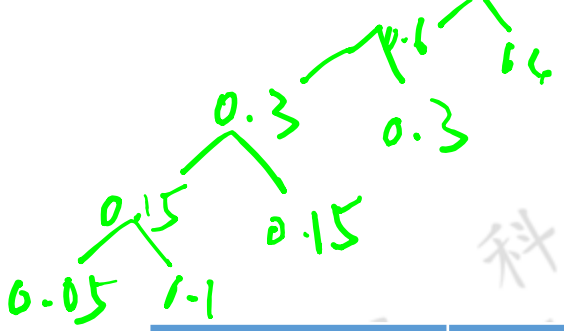
NeuHub AI Links All

参数量化（quantization）

- 减小参数精度
- 参数 clustering
- Huffman coding: 依据字符出现概率来进行不等长编码



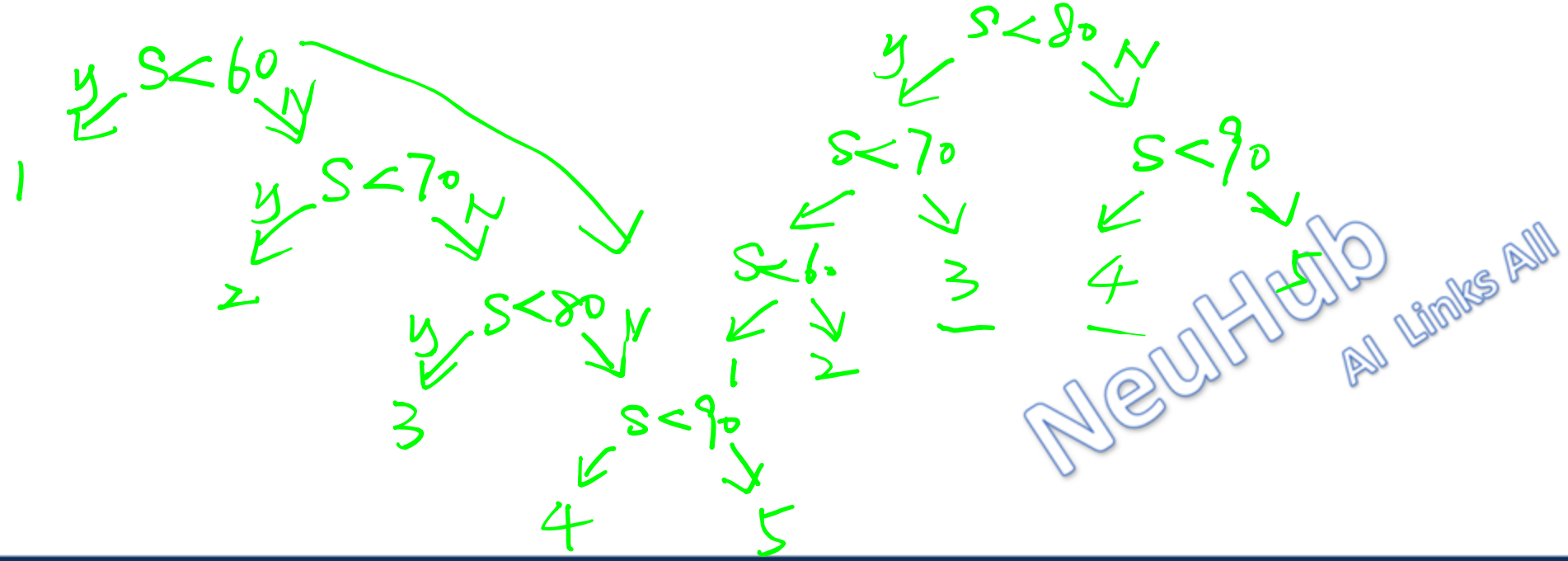
NeuHub
AI Links All



Huffman coding

分数段	0-59	60-69	70-79	80-89	90-100
五分制	1	2	3	4	5
比例	0.05	0.15	0.40	0.30	0.10

将百分制成绩转化为五分制成绩



NeuHub
AI Links All

Huffman coding

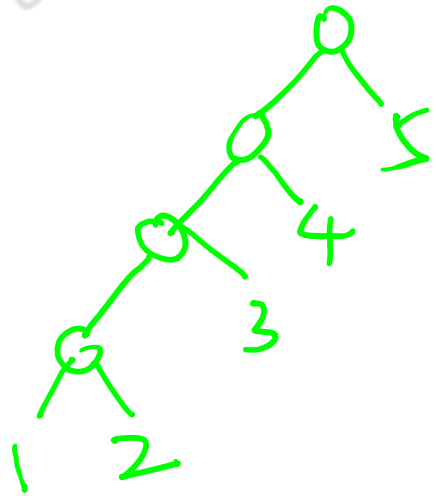
设二叉树有n个叶子结点，每个叶子结点带有权值Wk，从根结点到每个叶子结点的长度为lk，则每个叶子结点的带权路径长度之和（WPL）最小的时候，便称为最优二叉树或者哈夫曼树。

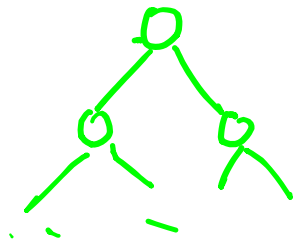
$$WPL = \sum W_k l_k$$

NeuHub
AI Links All

Huffman coding

假设有5个叶子节点，权值为{1,2,3,4,5}，用此权值序列可以构造出形状不同的多个二叉树。





WPL=5×1+4×2+3×3

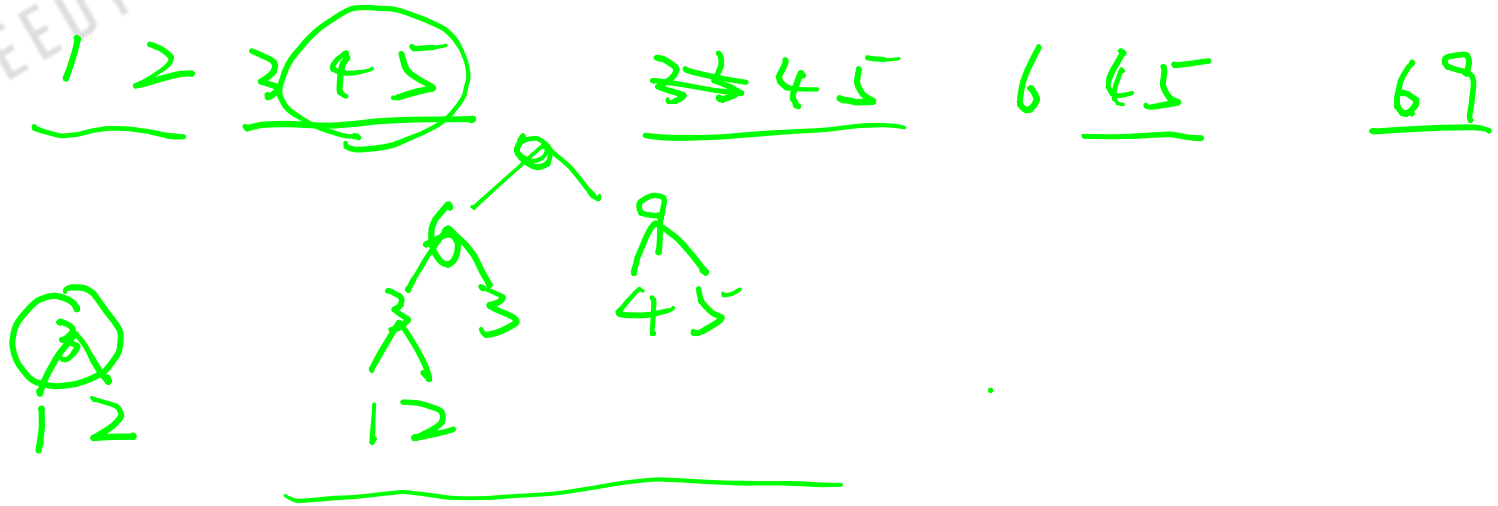
2×4+1×4=34

WPL

NeuHub
AI Links All

Huffman coding

哈夫曼树的构造：每次把权值最小的两棵二叉树合并



NeuHub
AI Links All

Huffman coding

给定一段字符串，如何进行编码，使得该字符串的编码存储空间最少？

例：aaaxuaxz

不等长编码：出现频率高的字符的编码短些，出现频率低的字符可以编码长些。

a : 1
x : 0
u : 11
z : 10

NeuHub
AI Links All

Huffman coding

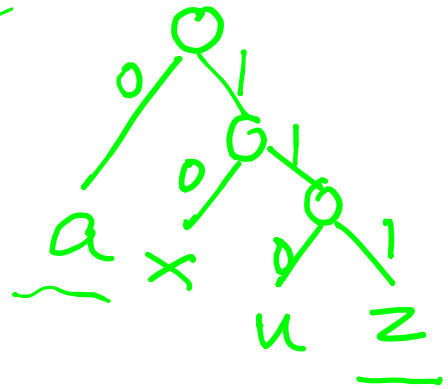
给定一段字符串，如何进行编码，使得该字符串的编码存储空间最少？
例：aaaxuaxz
不等长编码：出现频率高的字符的编码短些，出现频率低的字符可以编码长些。

如何避免二义性？任何字符的编码都不是其它字符的前缀。

NeuHub
AI Links All

Huffman coding

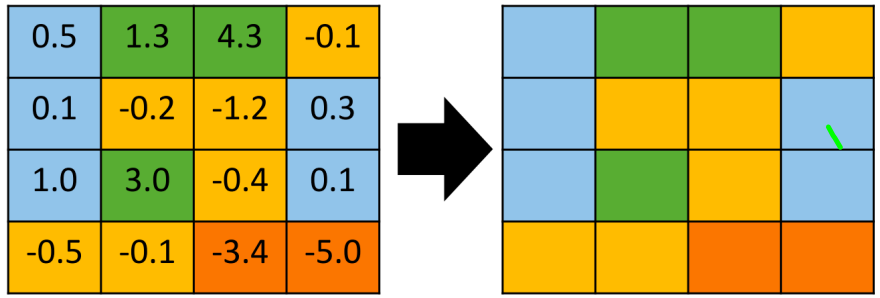
- 用二叉树进行编码：左右分支为0、1；字符只有叶子结点上
- 要想存储空间最小，就要去构造一棵哈夫曼树



NeuHub
AI Links All

参数量化（quantization）

- 减小参数精度
- 参数 clustering
- Huffman coding: 依据字符出现概率来进行不等长编码
- BinaryNet（有正则化效果）



NeuHub
AI Links All

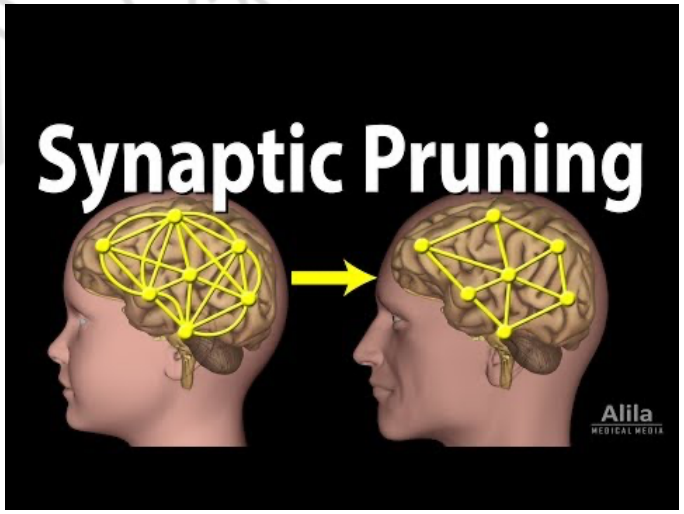
贪心科技 GREEDY TECHNOLOGY

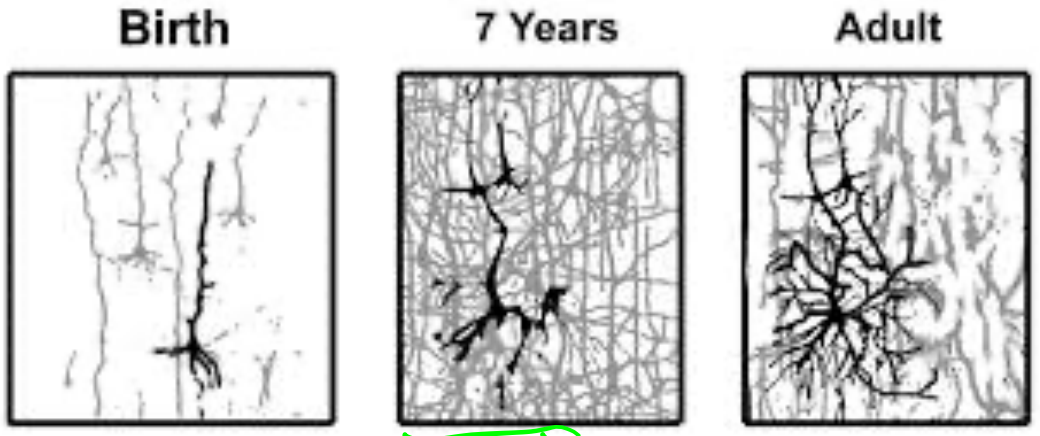
京东智联云 & 贪心科技 GREEDY TECHNOLOGY

模型裁剪/剪枝（Pruning）

NeuHub AI Links All

模型裁剪/剪枝（Pruning）





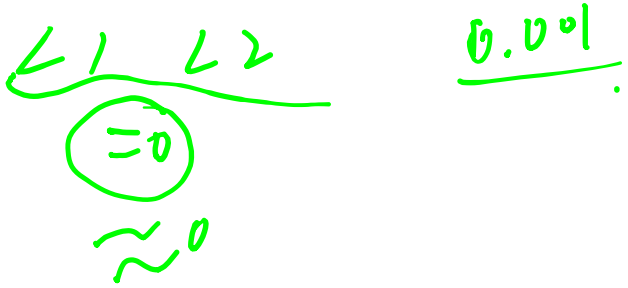
发生在人类身上的突触剪枝

NeuHe
AI Link All

模型裁剪/剪枝（Pruning）

挑选出模型中不重要的参数，将其剔除而不会对模型的效果造成太大的影响。
在剔除不重要的参数之后，通过一个retrain的过程来恢复模型的性能。

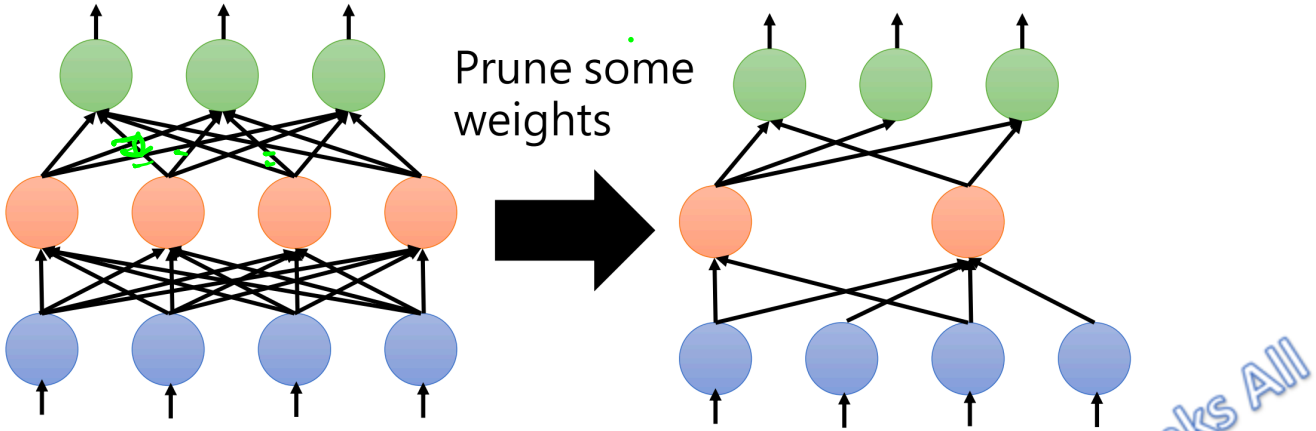
- 单个权重(Weight)剪枝
- 单个神经元(Neuro)剪枝



NeuHub
AI Links All

模型裁剪/剪枝（Pruning）

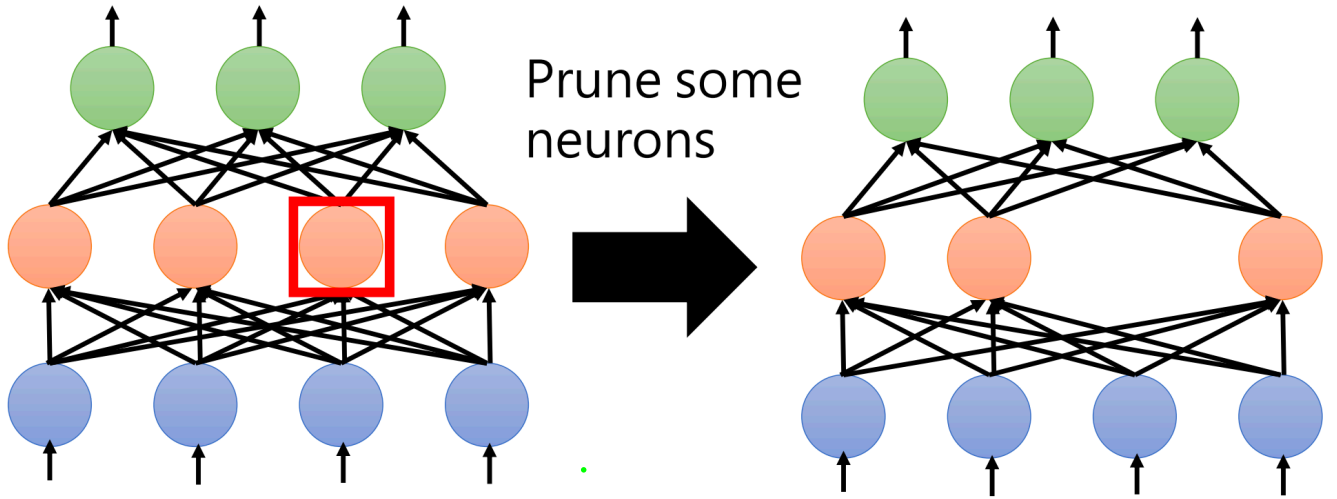
单个权重(Weight)剪枝：
导致网络连接不规整



AI Links All

模型裁剪/剪枝（Pruning）

单个神经元(Neuro)剪枝

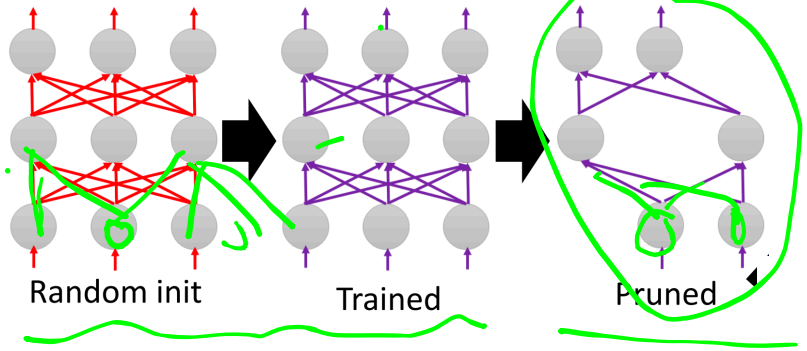


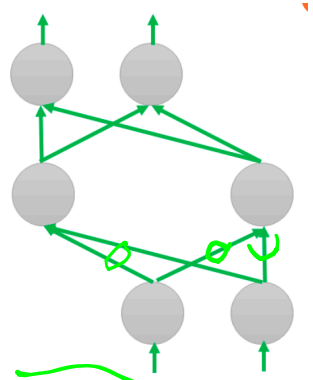
模型裁剪/剪枝（Pruning）

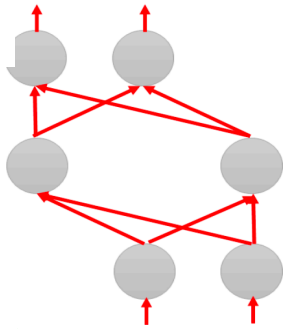
为什么不直接训练小模型呢？
ICLR 2019最佳论文《The Lottery Ticket Hypothesis》

1. 作者通过剪枝将模型参数减少到10%-20%

2. 剪枝后的网络使用原初始化参数初始化后重新训练会取得原模型类似的性能，但在随机重初始化后重新训练会表现的更糟糕。







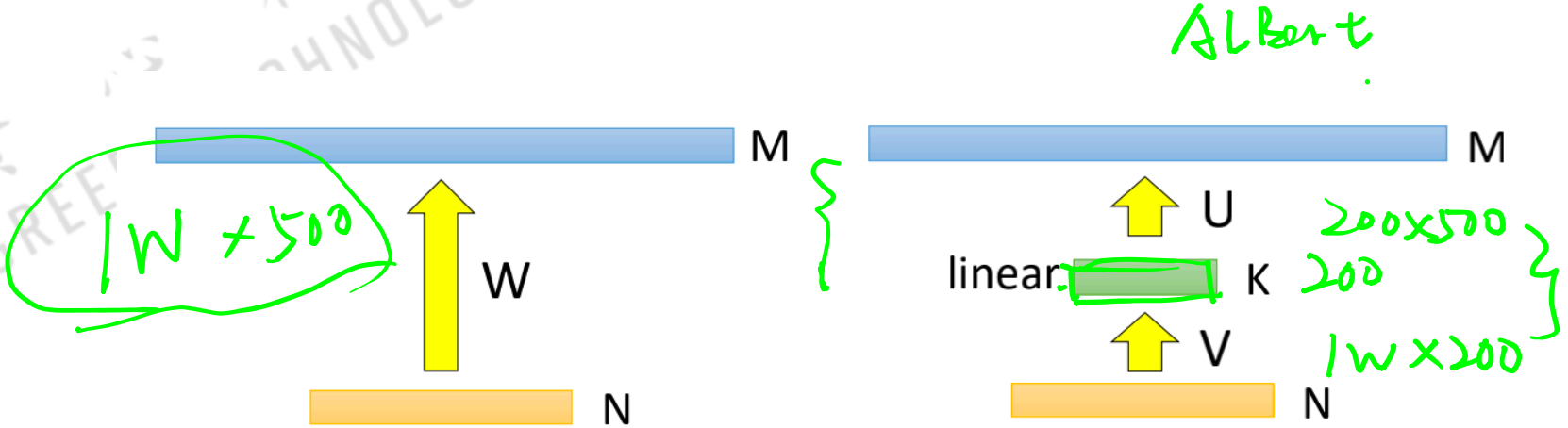
贪心科技 GREEDY TECHNOLOGY

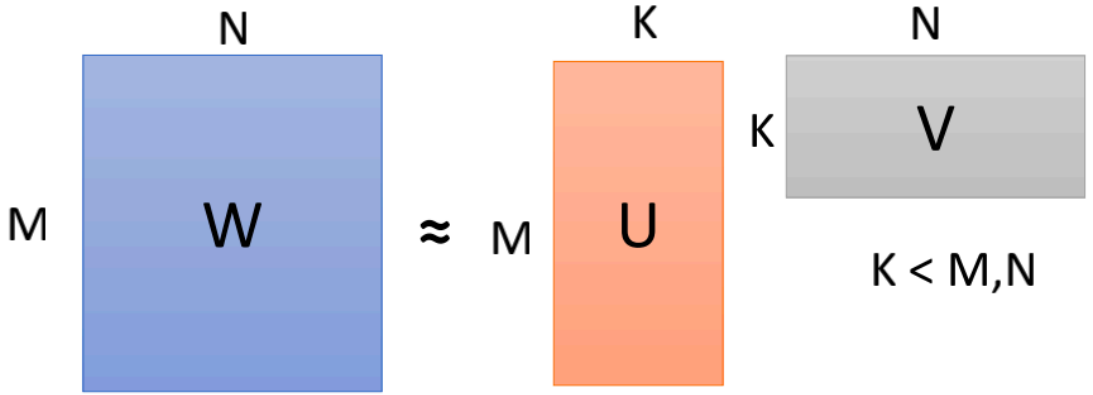
京东智联云 & 贪心科技 GREEDY TECHNOLOGY

模型结构重构

NeuHub AI Links All

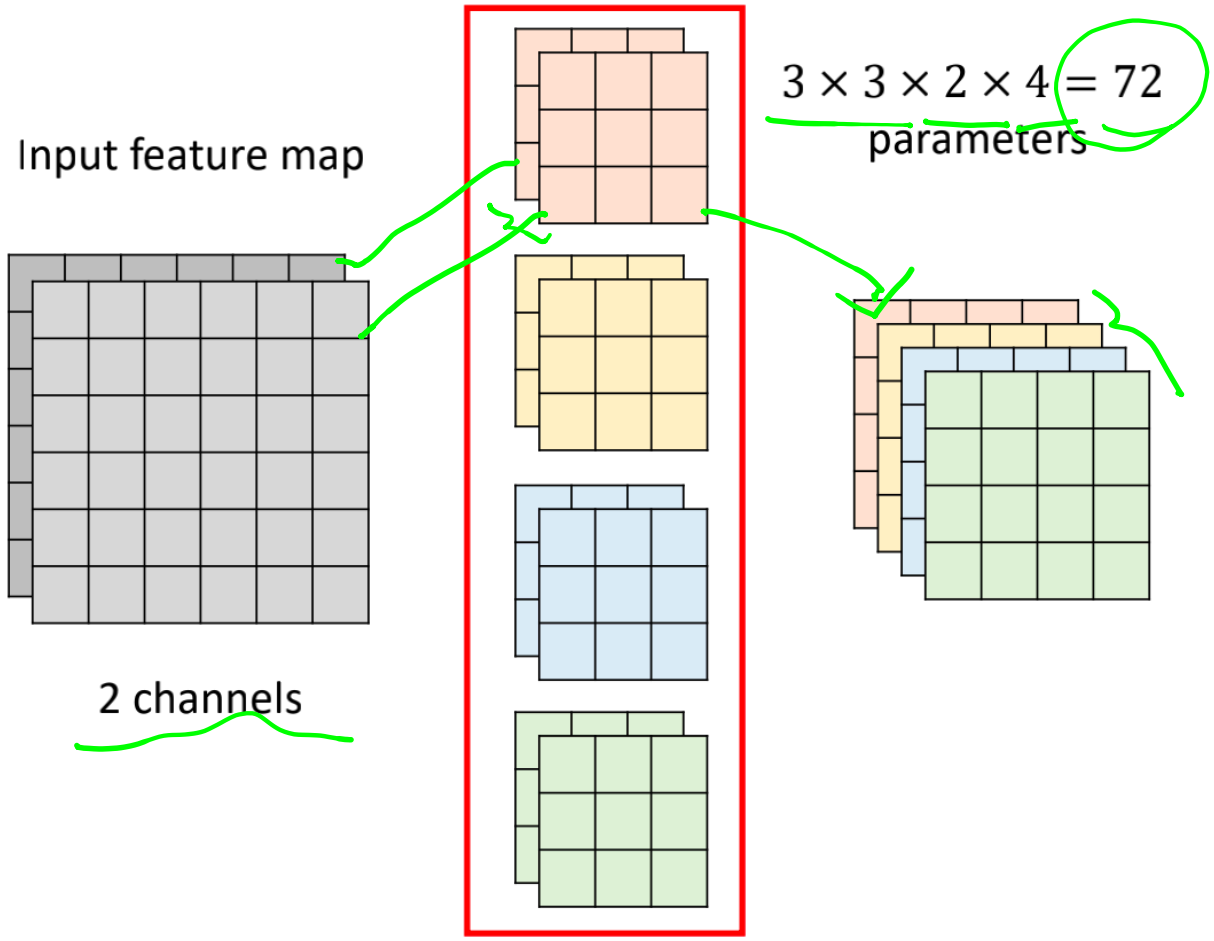
模型结构重构





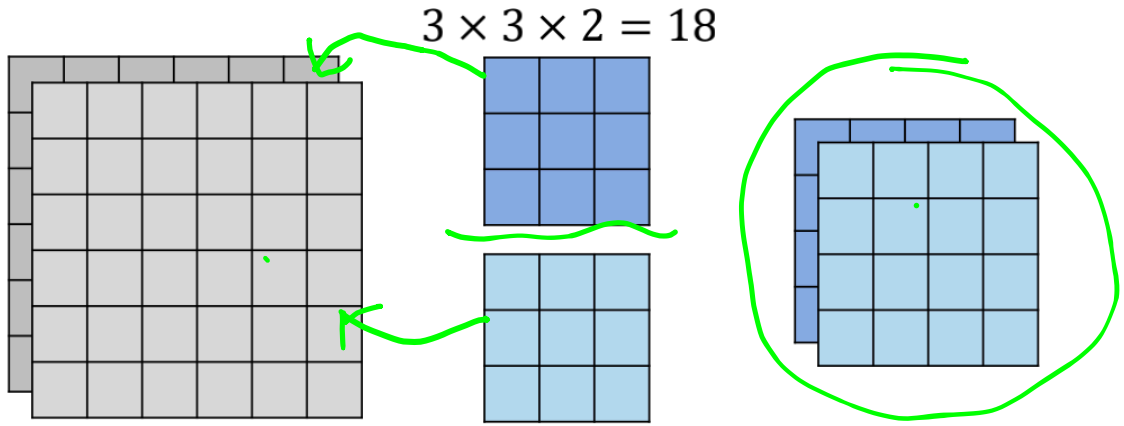
AI Links All

模型结构重构

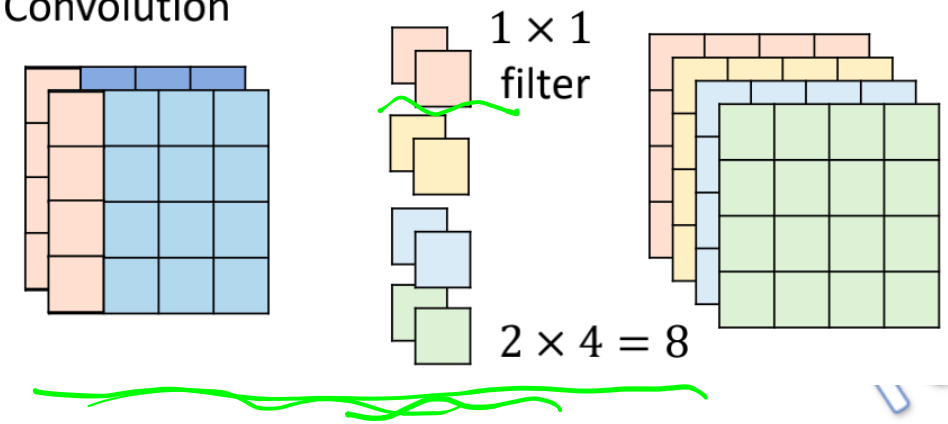


模型结构重构

1. Depthwise Convolution



2. Pointwise Convolution



京东智联云 & GREEDY TECHNOLOGY
AI Links All

贪心科技 GREEDY TECHNOLOGY

京东智联云 & 贪心科技 GREEDY TECHNOLOGY

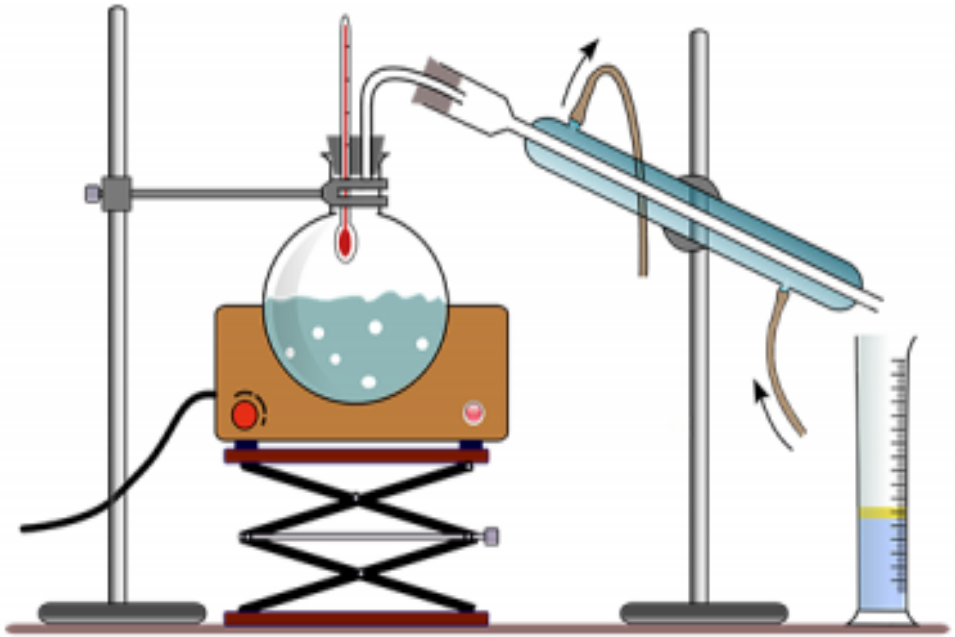
知识蒸馏

NeuHub AI Links All

知识蒸馏

KD

- 源于2015年Hinton发表的一篇论文:
Distilling the Knowledge in a Neural Network

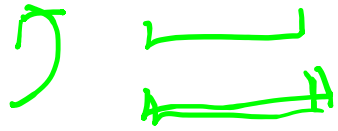


知识蒸馏

Teacher Model和Student Model

- 2个阶段:

- 原始模型训练
- 精简模型训练



注：在Hinton的论文中，作者将问题限定在分类问题下，或者其他本质上属于分类问题的问题，该类问题的共同点是模型最后会有一个softmax层，其输出值对应了相应类别的概率值。

NeuHub AI Links All

知识蒸馏

为什么要学习softmax层的输出概率分布？

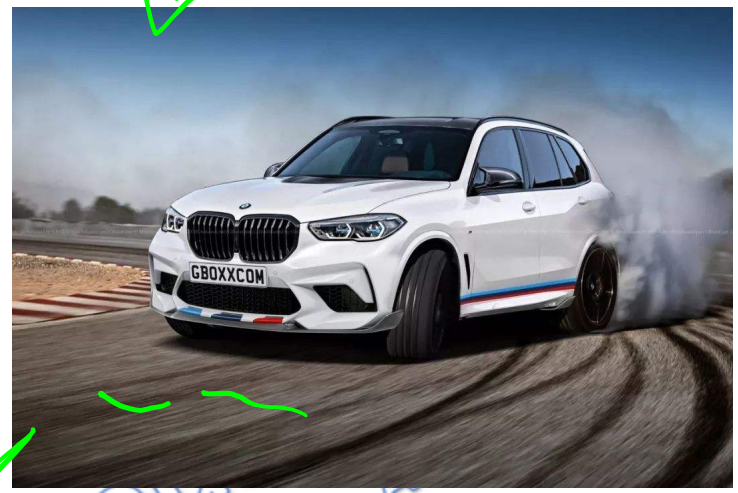
机器学习最根本的目的：在于训练出在某个问题上泛化能力强的模型。

一句话总结KD：
传统training过程(hard targets): 对ground truth求极大似然
KD的training过程(soft targets): 用large model的class probabilities作为soft targets

NeuHub
AI Links All

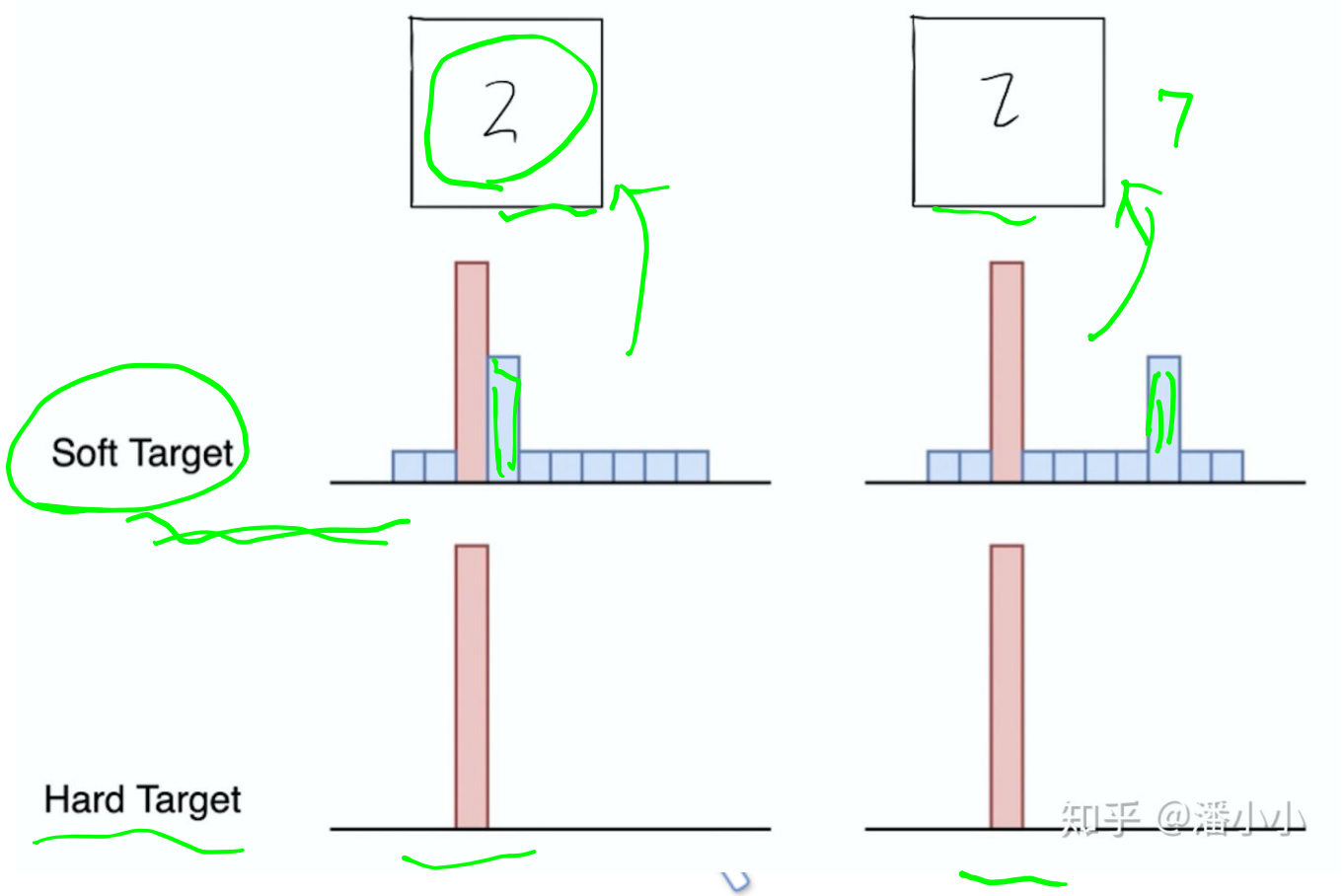
知识蒸馏

为什么学习softmax层的输出概率分布，就能学习泛化能力？



知识蒸馏

为什么学习softmax层的输出概率分布，就能学习泛化能力？



知识蒸馏-温度T

$$y_i = \frac{exp(x_i)}{\sum_j exp(x_j)} \Rightarrow y_i = \frac{exp(x_i/T)}{\sum_j exp(x_j/T)}$$

$T = 100$

logits

$x_1 = 100$

$x_2 = 10$

$x_3 = 1$

one-hot

$y_1 = 1$

$y_2 \approx 0$

$y_3 \approx 0$

$x_1/T = 1$

$x_2/T = 0.1$

$x_3/T = 0.01$

$y_1 = 0.56$

$y_2 = 0.23$

$y_3 = 0.21$

贪心科技 | 让每个人享受个性化教育服务

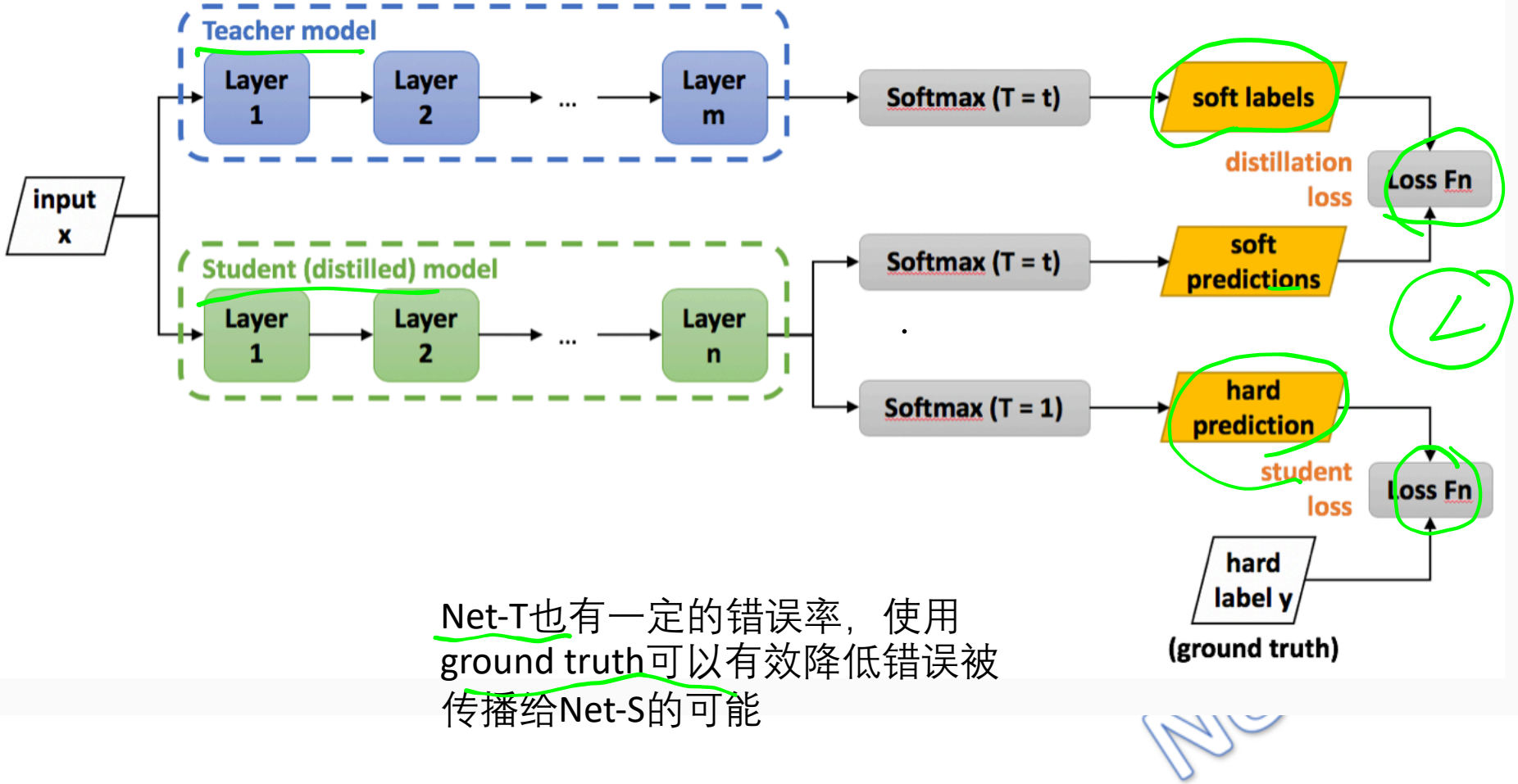
知识蒸馏-温度

温度T的特点

- 原始的softmax函数是 T = 1时的特例。
- 温度越高，softmax上各个值的分布就越平均。
- 不管温度T怎么取值，Soft target都有忽略小的p所携带信息的倾向。

NeuHub
AI Links All

知识蒸馏-损失函数

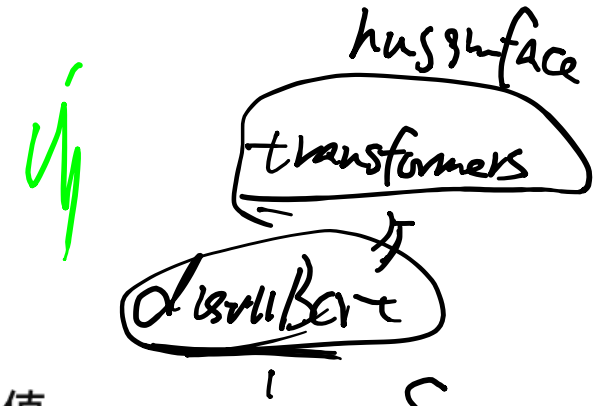


Net-T也有一定的错误率，使用ground truth可以有效降低错误被传播给Net-S的可能

知识蒸馏-损失函数

京东智联云 & AI GREEDY TECHNOLOGY

$$L = \alpha L_{soft} + \beta L_{hard}$$



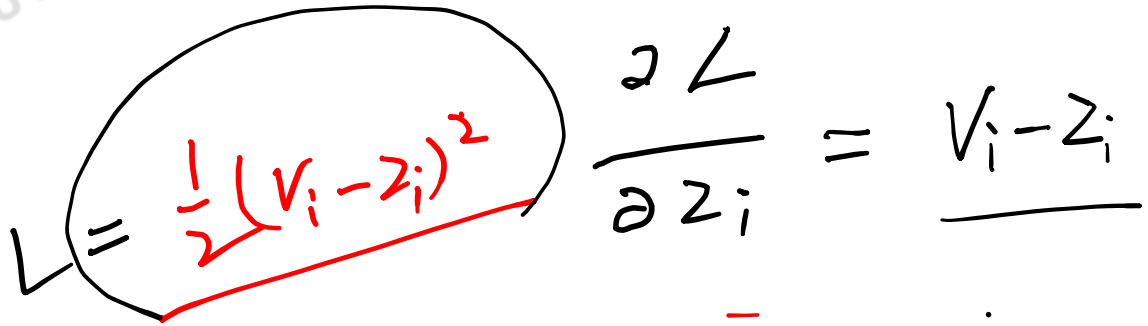
- v_i : Net-T的logits
- z_i : Net-S的logits
- p_i^T : Net-T的在温度=T下的softmax输出在第i类上的值
- q_i^T : Net-S的在温度=T下的softmax输出在第i类上的值
- c_i : 在第i类上的ground truth值，$c_i \in \{0, 1\}$, 正标签取1，负标签取0.
- N : 总标签数量

$$L_{soft} = -\sum^{N} p_j^T \log(q_j^T)$$

$$L_{hard} = -\sum^{N} c_j \log(q_j^1)$$

NeuHub AI Links All

知识蒸馏-损失函数



直接match logits，用mse作为损失函数行不行？

NeuHub
AI Links All

知识蒸馏-温度

京东智联云 & AI GREEDY TECHNOLOGY

$$L = \alpha L_{soft} + \beta L_{hard}$$

- v_i : Net-T的logits
- z_i : Net-S的logits
- p_i^T : Net-T的在温度=T下的softmax输出在第i类上的值
- q_i^T : Net-S的在温度=T下的softmax输出在第i类上的值
- c_i : 在第i类上的ground truth值，$c_i \in \{0,1\}$, 正标签取1，负标签取0.
- N : 总标签数量

① $T \to \infty$

$1 + x/T \approx e^{x/T}$

$$\approx \frac{1}{T}\left(\frac{1+z_i/T}{N+\sum_j z_j/T} - \frac{1+v_i/T}{N+\sum_j v_j/T}\right)$$

$\sum_j z_j = 0$ $\sum_j v_j = 0$

$$\frac{\partial L_{soft}}{\partial z_i} \approx \frac{1}{NT^2}(z_i - v_i)$$

② logits

$T \to \infty$

x → softmax → cross entropy

$$\frac{\partial L_{soft}}{\partial z_i} = \frac{1}{T}(q_i - p_i)$$

$$= \frac{1}{T}\left(\frac{e^{z_i/T}}{\sum_j e^{z_j/T}} - \frac{e^{v_i/T}}{\sum_j e^{v_j/T}}\right)$$

NeuHub AI Links All

贪心科技 | 让每个人享受个性化教育服务

知识蒸馏-总结

Why soft target？

- 用T model的class probabilities作为soft targets，让S model直接学习到其泛化能力。
- soft targets比hard targets带有更多的信息。

温度的高低改变的是Net-S训练过程中对负标签的关注程度

取舍:

- 从有部分信息量的负标签中学习 --> 温度要高一些
- 防止受负标签中噪声的影响 -->温度要低一些

T的选择和Net-S的大小有关

NeuHub
AI Links All

贪心科技 GREEDY TECHNOLOGY

京东智联云 & 贪心科技 GREEDY TECHNOLOGY

14 (5)

more about 知识蒸馏

NeuHub AI Links All

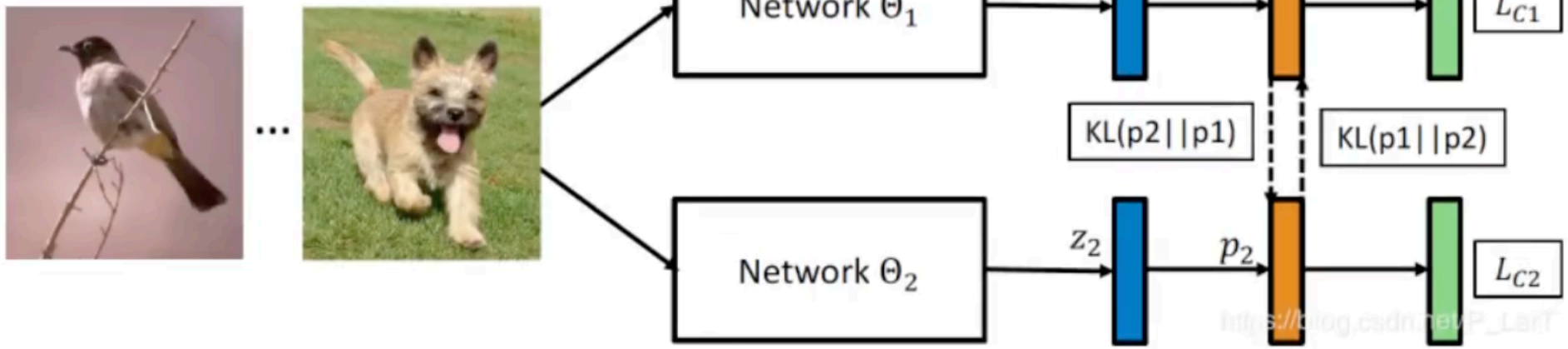
Deep Mutual Learning

- 相互学习策略(deep mutual learning, DML): 打破预先定义好的“强弱关系”
- 作者通过在CIFAR-100和Market-1501数据集上的实验，表明DML网络在分类和任务重识别任务中的有效性。
- DML的成功揭示了没有强大的教师网络是可行的，相互学习的对象是由一个个简单的学生网络组成的集合。

.

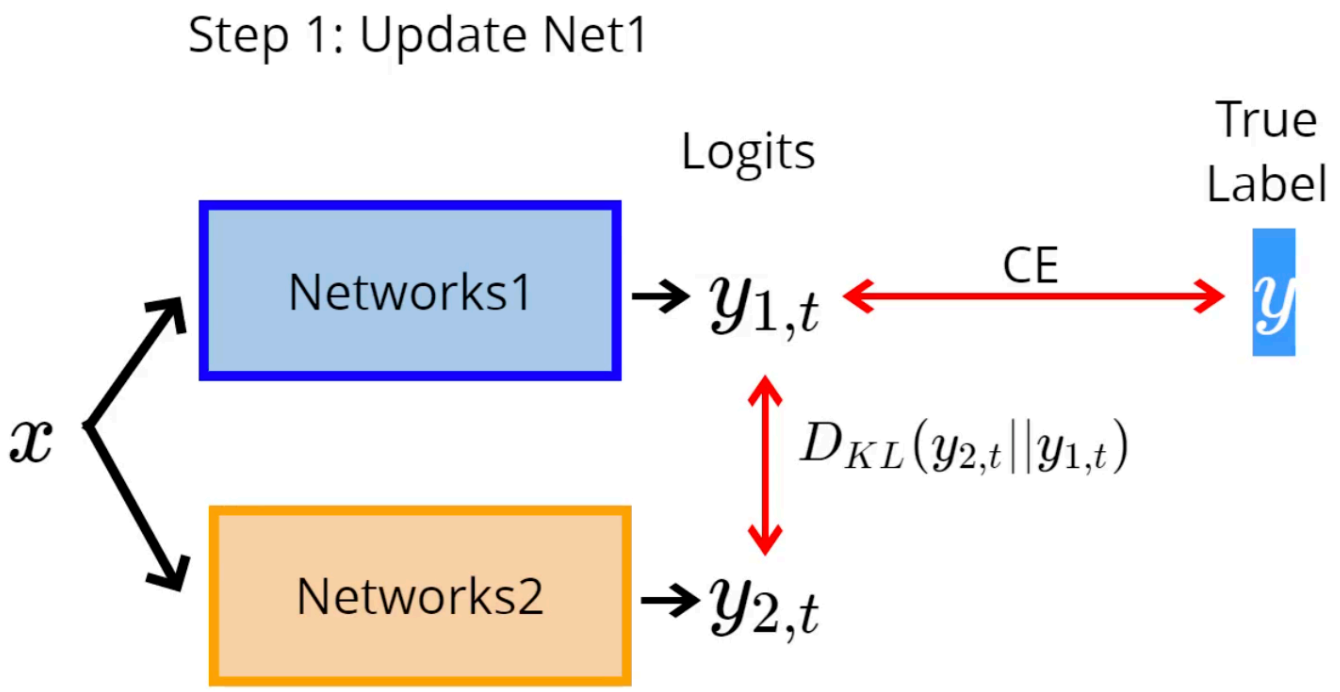
NeuHub
AI Links All

Deep Mutual Learning



贪心科技 | 让每个人享受个性化教育服务

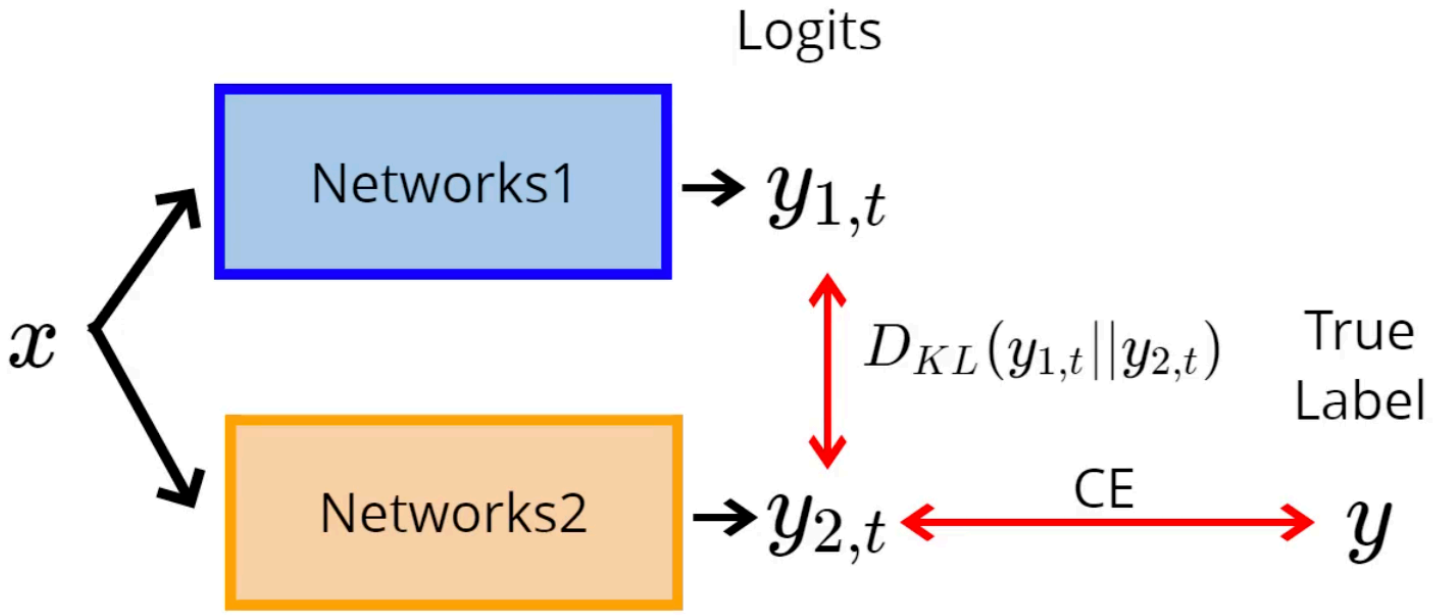
Deep Mutual Learning



$$Loss_1 = D_{KL}(y_{2,t}||y_{1,t}) + \text{CrossEntropy}(y, y_{1,t})$$

Deep Mutual Learning

Step 2: Update Net2



$$Loss_2 = D_{KL}(y_{1,t}||y_{2,t}) + \text{CrossEntropy}(y, y_{2,t})$$

Deep Mutual Learning

贪心科技 GREEDY TECHNOLOGY

- 在训练过程中，每个学生网络的损失函数由两部分组成：
  - 传统的监督学习损失
  - 模仿损失，使每个学生的预测类别与其他学生的类别预测概率保持一致。

- 实验证明，在这种基于同伴教学（peer-teaching）的训练中，每个学生网络的学习效果都比在传统的监督学习场景中单独学习要好得多。

- But why ?

NeuHub AI Links All

Deep Mutual Learning

- 每个网络从不同的初始条件开始
- 网络效果随队列中网络的数量增加而增加
- 适用于异构群组
- 相互训练的大型网络也能提高性能
- 整个队列也可以整合为一个高效的集成模型。

NeuHub
AI Links All

FitNet(Hints for Thin Deep Nets)

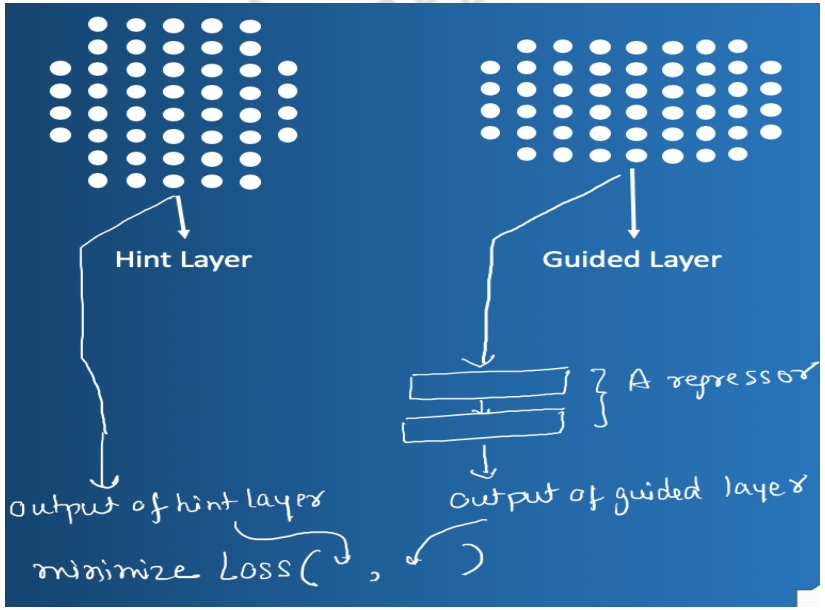
给点提示

Teacher Net

Student Net

NeuHub AI Links All

FitNet(Hints for Thin Deep Nets)



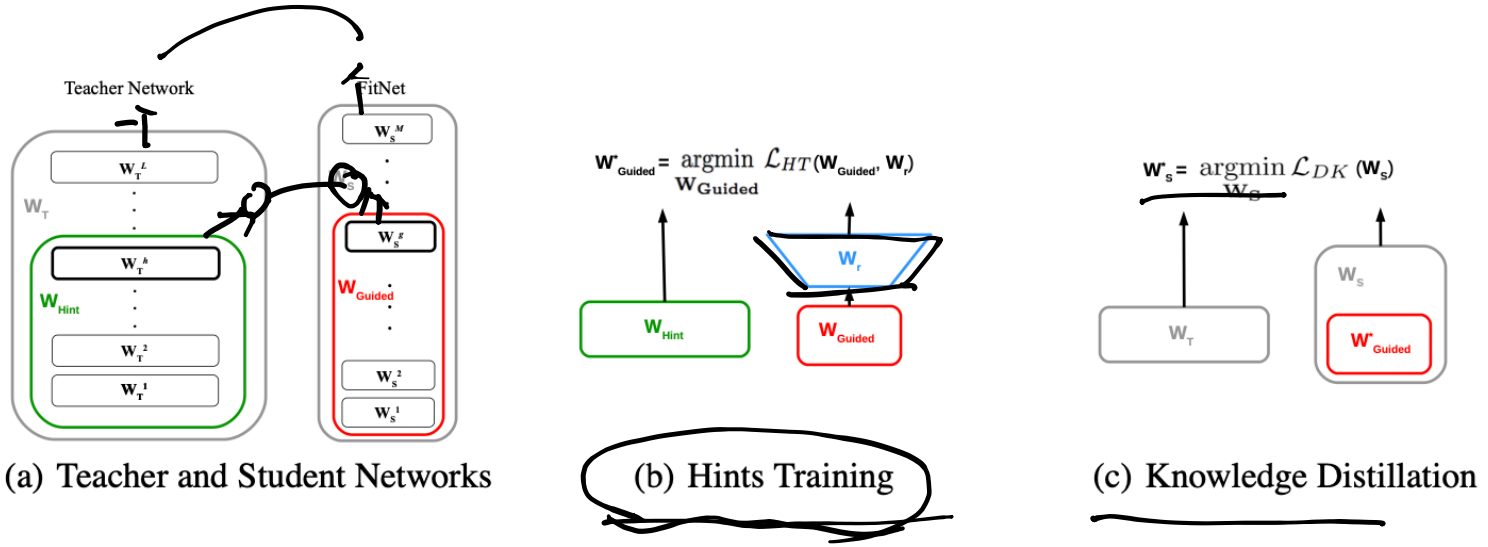
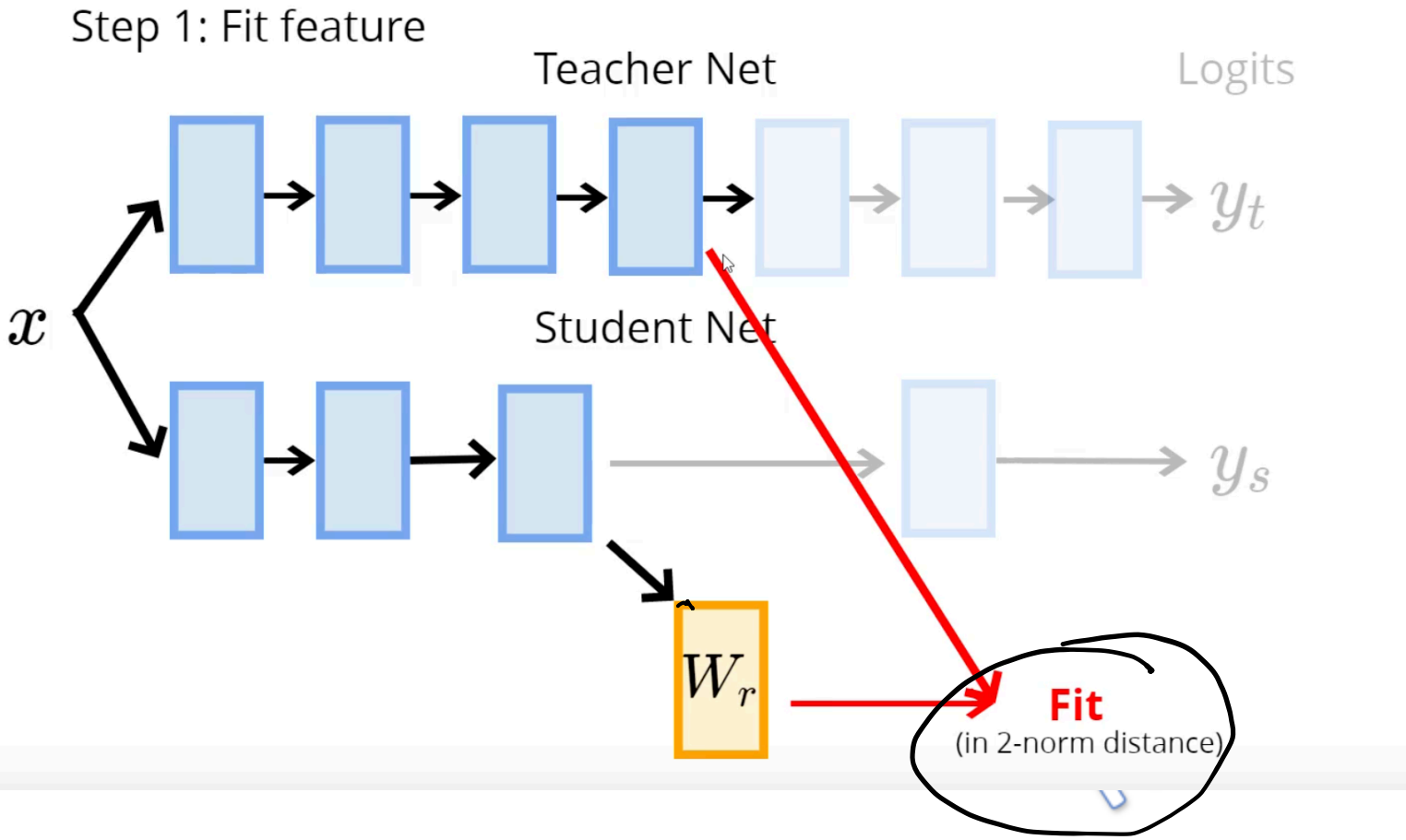


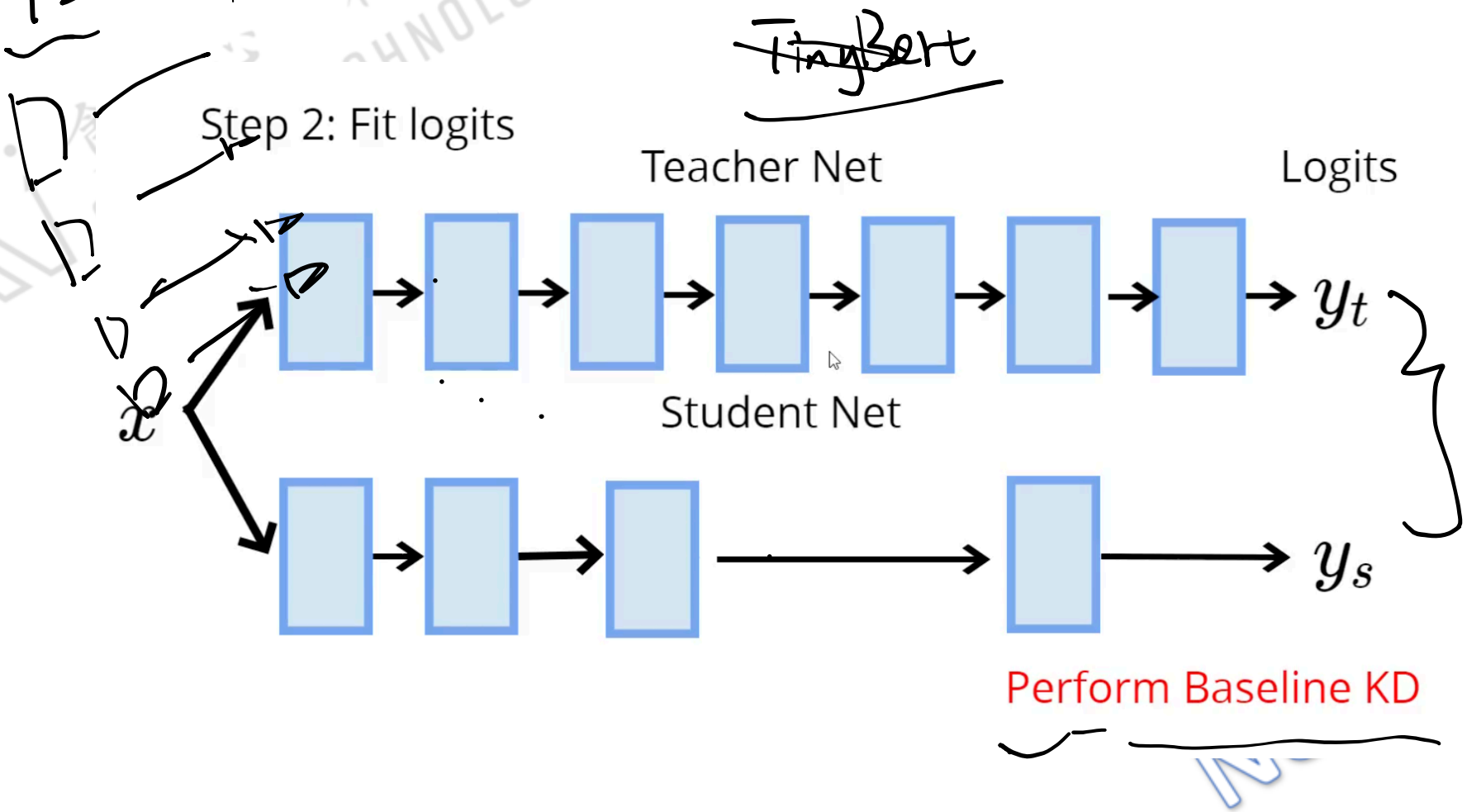
Figure 1: Training a student network using hints.

先让student学习teacher如何产生内部feature，再应用baseline KD

FitNet(Hints for Thin Deep Nets)



FitNet(Hints for Thin Deep Nets)



FitNet(Hints for Thin Deep Nets)

- 第一个阶段利用Hint-based loss诱导学生网络达到一个合适的初始化状态（只更新W_Guided与W_r）
- 第二个阶段利用教师网络的soft label指导整个学生网络的训练（即KD）
- soft target相关部分所占比重逐渐降低

Algorithm 1 FitNet Stage-Wise Training.
The algorithm receives as input the trained parameters $\mathbf{W_T}$ of a teacher, the randomly initialized parameters $\mathbf{W_S}$ of a FitNet, and two indices h and g corresponding to hint/guided layers, respectively. Let $\mathbf{W_{Hint}}$ be the teacher's parameters up to the hint layer h. Let $\mathbf{W_{Guided}}$ be the FitNet's parameters up to the guided layer g. Let $\mathbf{W_r}$ be the regressor's parameters. The first stage consists in pre-training the student network up to the guided layer, based on the prediction error of the teacher's hint layer (line 4). The second stage is a KD training of the whole network (line 6).

Input: $\mathbf{W_S}, \mathbf{W_T}, g, h$
Output: $\mathbf{W_S^*}$
1: $\mathbf{W_{Hint}} \leftarrow \{\mathbf{W_T}^1, \ldots, \mathbf{W_T}^h\}$
2: $\mathbf{W_{Guided}} \leftarrow \{\mathbf{W_S}^1, \ldots, \mathbf{W_S}^g\}$
3: Intialize $\mathbf{W_r}$ to small random values
4: $\mathbf{W^*_{Guided}} \leftarrow \underset{\mathbf{W_{Guided}}}{\text{argmin}}\ \mathcal{L}_{HT}(\mathbf{W_{Guided}}, \mathbf{W_r})$
5: $\{\mathbf{W_S}^1, \ldots, \mathbf{W_S}^g\} \leftarrow \{\mathbf{W_{Guided}}^{*1}, \ldots, \mathbf{W_{Guided}}^{*g}\}$
6: $\mathbf{W_S^*} \leftarrow \underset{\mathbf{W_S}}{\text{argmin}}\ \mathcal{L}_{KD}(\mathbf{W_S})$

TinyBERT：模型小7倍，速度快8倍

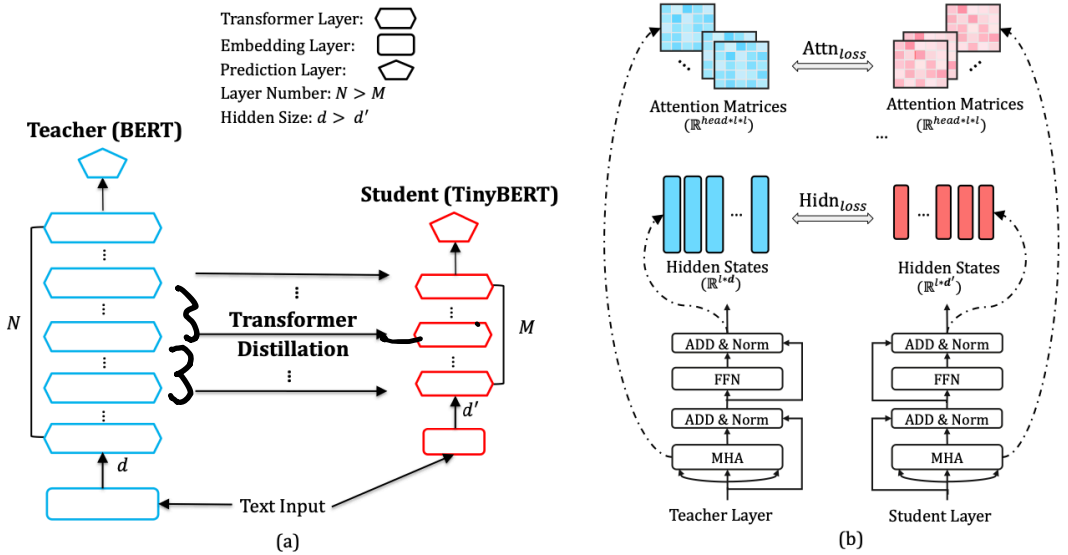


Figure 1: An overview of Transformer distillation: (a) the framework of Transformer distillation, (b) the details of Transformer-layer distillation consisting of Attn_{loss}(attention based distillation) and Hidn_{loss}(hidden states based distillation).

$$\mathcal{L}_{\text{layer}}(S_m, T_{g(m)}) = \begin{cases} \mathcal{L}_{\text{embd}}(S_0, T_0), & m = 0 \\ \mathcal{L}_{\text{hidn}}(S_m, T_{g(m)}) + \mathcal{L}_{\text{attn}}(S_m, T_{g(m)}), & M \geq m > 0 \\ \mathcal{L}_{\text{pred}}(S_{M+1}, T_{N+1}), & m = M+1 \end{cases}$$

Attention to Attention Transfer

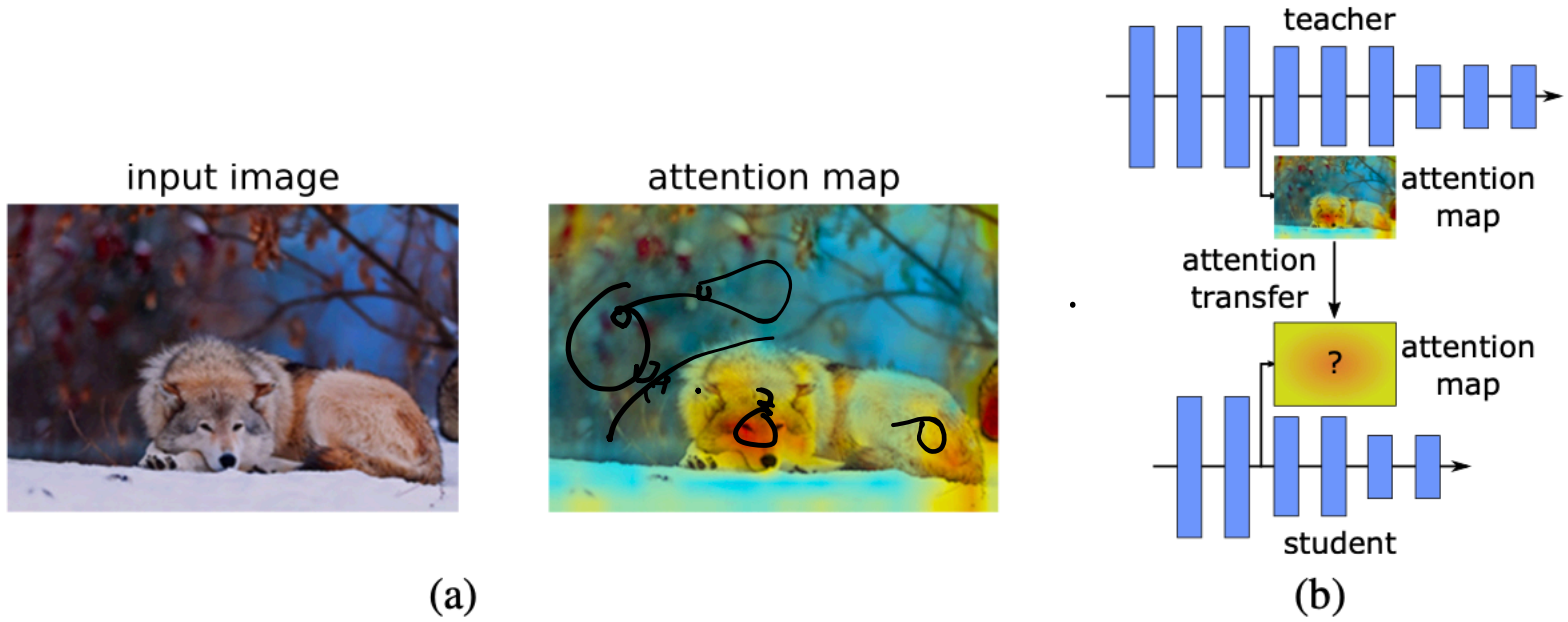
给点提示也不行？
划重点

Teacher Net

Student Net

NeuHub AI Links All

Attention to Attention Transfer



How attention differs within artificial vision systems, and can we use attention information in order to improve the performance of convolutional neural networks ? More specifically, can a teacher network improve the performance of another student network by providing to it information about where it looks, i.e., about where it concentrates its attention into ?

Attention to Attention Transfer

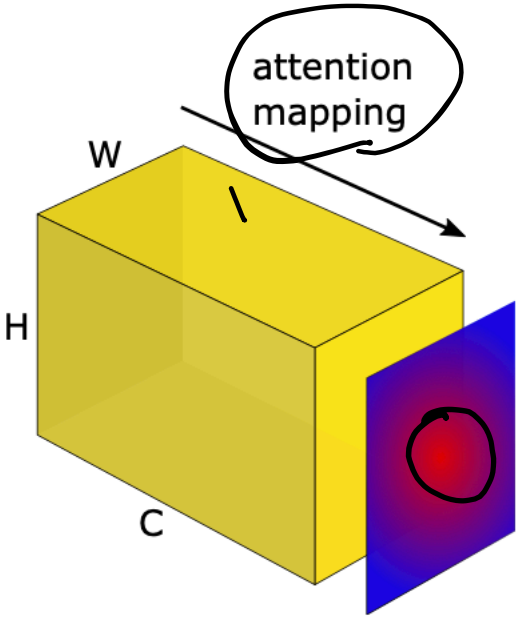


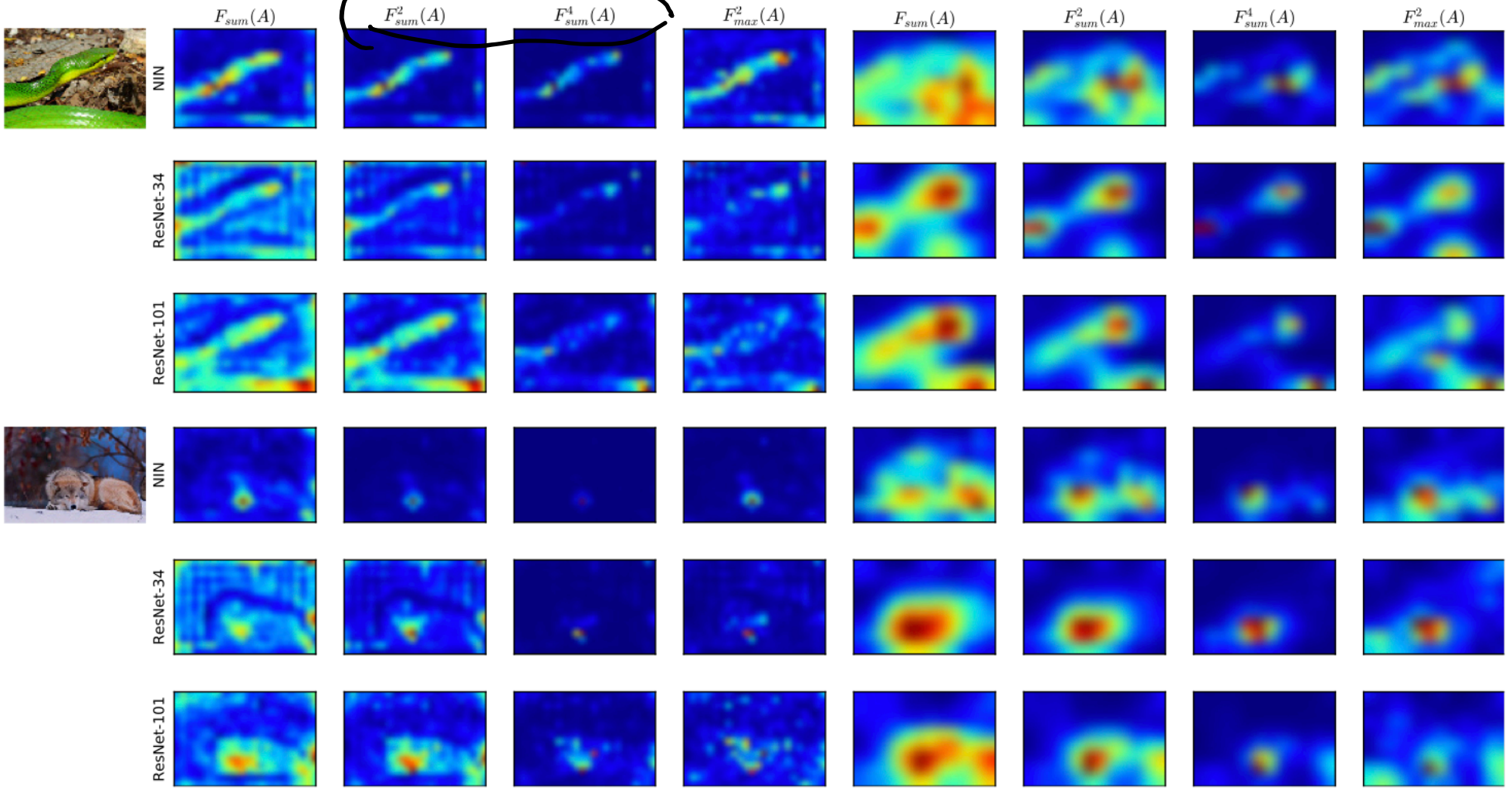
Figure 3: Attention mapping over feature dimension.

- sum of absolute values: $F_{\text{sum}}(A) = \sum_{i=1}^{C} |A_i|$
- sum of absolute values raised to the power of p (where $p > 1$): $F_{\text{sum}}^{p}(A) = \sum_{i=1}^{C} |A_i|^p$
- max of absolute values raised to the power of p (where $p > 1$): $F_{\text{max}}^{p}(A) = \max_{i=1,C} |A_i|^p$

- 各个channel相同位置绝对值求和
- 各个channel相同位置p次方求和：对比1，会更加注重于响应高的地方
- 各个channel相同位置p次方求最大值

NeuHub AI Links All

Attention to Attention Transfer



贪心科技 | 让每个人享受个性化教育服务

Attention to Attention Transfer

$$\mathcal{L}_{AT} = \mathcal{L}(\mathbf{W}_S, x) + \frac{\beta}{2} \sum_{j \in \mathcal{I}} \left\| \frac{Q_S^j}{\|Q_S^j\|_2} - \frac{Q_T^j}{\|Q_T^j\|_2} \right\|_p ,$$

attention mapping function	error
no attention transfer	8.77
F_{sum}	7.99
F_{sum}^2	7.93
F_{sum}^4	8.09
F_{max}^1	8.08

第二种方法效果最好

- Qs Qt为第j对student和teacher的attention map
- p取2
- beta取1000

NeuHub
AI Links All

Knowledge Distillation with Adversarial Samples Supporting Decision Boundary

Teacher Net

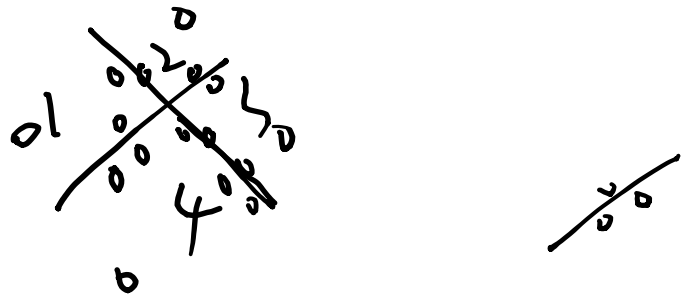
怎么

给点提示也不行？
划重点样本

Student Net

NeuHub AI Links All

Knowledge Distillation with Adversarial Samples Supporting Decision Boundary



In general, samples near the decision boundary of a classifier have a larger impact on the performance than those far apart from it (Cortes and Vapnik 1995). Therefore, if we can generate samples close to the decision boundary, the knowledge of a teacher network would be transferred more effectively by utilizing those samples.

NeuHub
AI Links All

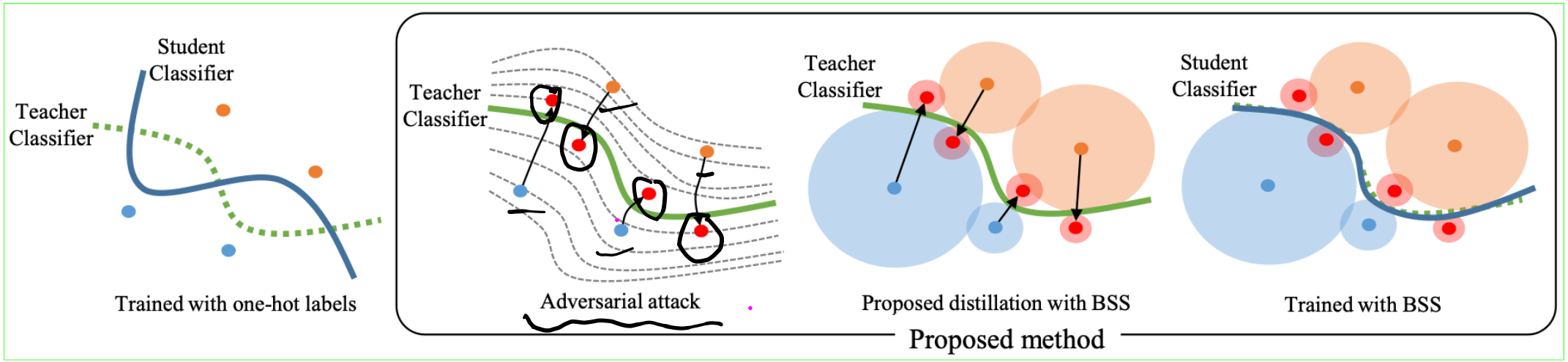
Knowledge Distillation with Adversarial Samples Supporting Decision Boundary

To obtain the informative samples close to the decision boundary, we utilize an adversarial attack (Goodfellow, Shlens, and Szegedy 2015; Moosavi-Dezfooli, Fawzi, and Frossard 2016).
An adversarial attack tries to find a small modification that can change the class of a sample, i.e., it tries to move the sample beyond a nearby decision boundary.

对抗攻击，adversarial attack：对输入样本故意添加一些人无法察觉的细微的干扰，导致模型以高置信度给出一个错误的输出。
例：使得DNN判别图片类型时，误判的同时，图片的改变尽可能少。

NeuHa
AI Link

Knowledge Distillation with Adversarial Samples Supporting Decision Boundary



boundary supporting sample (BSS)

Knowledge Distillation with Adversarial Samples Supporting Decision Boundary

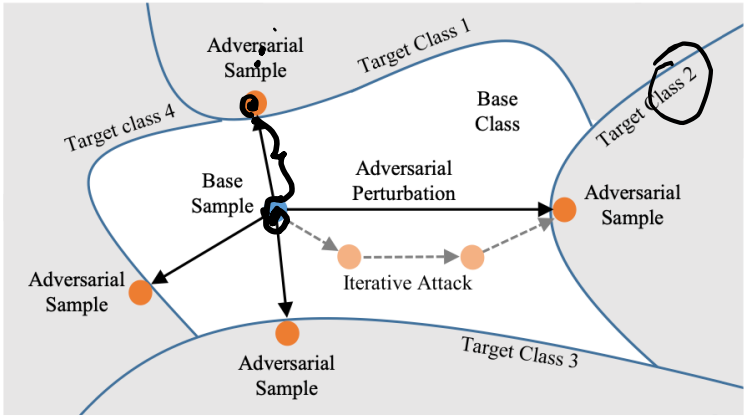


Figure 2: Iterative scheme to find BSSs for a base sample

b: 不原来类别　k: x后的目标类

$f_b(x)$: b对应的logit　　i: 第i次变化x

$$L(x) = f_b(x) - f_k(x)$$

$$\begin{cases} \max \text{存在} f_k(x) > \max(f_b(x), f_k(x)) \\ f_k(x) > f_b(x) \end{cases} \quad \begin{matrix} L(x) < 0 \\ L(x) \geq 0 \end{matrix}$$

NeuHub AI Links All

Knowledge Distillation with Adversarial Samples Supporting Decision Boundary

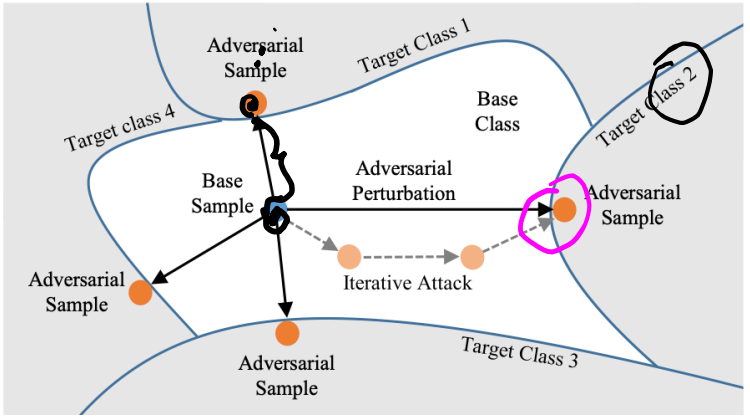


Figure 2: Iterative scheme to find BSSs for a base sample

b: 不原来类别　k: x后的目标类

$f_b(x)$: b对应的logit　　i: 第i次变化x

$L(x) = f_b(x) - f_k(x)$　　负梯度

$$\begin{cases} \max \\ 存在\ f_k(x) > \max(f_b(x), f_k(x)) \\ f_k(x) > f_b(x) \end{cases} \quad (x)$$

$L(x) < 0$

$L(x) \geq 0$ ✓

$$x_{i+1} = x_i - \eta(L(x_i)+\epsilon)\cdot \frac{\nabla L(x_i)}{\|\nabla L(x_i)\|_2}$$

$$L(x_{i+1}) \approx L(x_i) + \nabla L(x_i)(x_{i+1}-x_i)$$

$$L(x_{i+1}) \approx L(x_i) + \nabla L(x_i)\cdot\left(-\eta(L(x_i)+\epsilon)\frac{\nabla L(x_i)}{\|\ \|_2}\right)$$

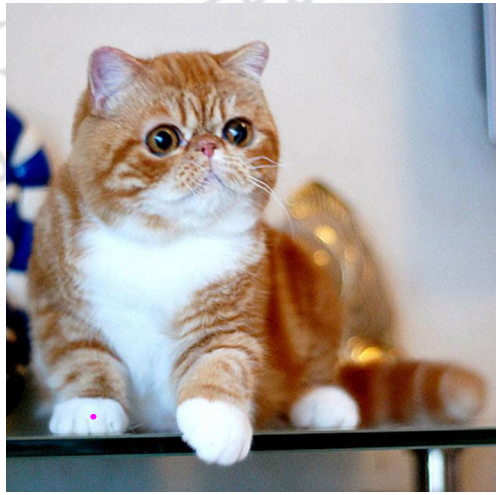
$$\approx L(x_i) + (-\eta\cdot L(x_i) - \eta\epsilon)\cdot\|\ \|_2$$

$$\approx L(x_i)(1-\eta\|\ \|_2) - \eta\epsilon\|\ \|_2$$

>0　　<0

HeuHub AI Links All

Relational KD

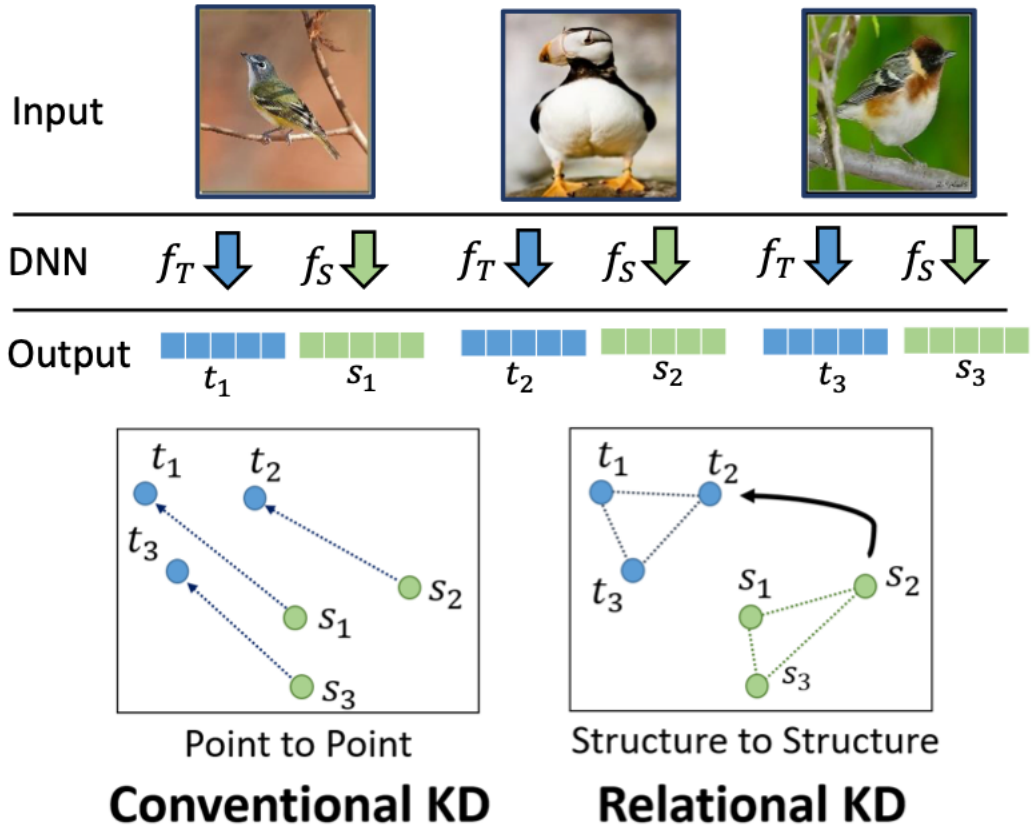


有毛，有四肢，喵喵叫...

有无毛

NeuHub
AI Links All

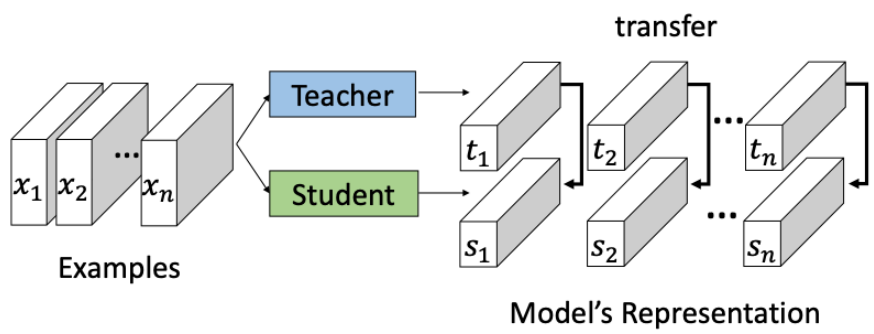
Relational KD



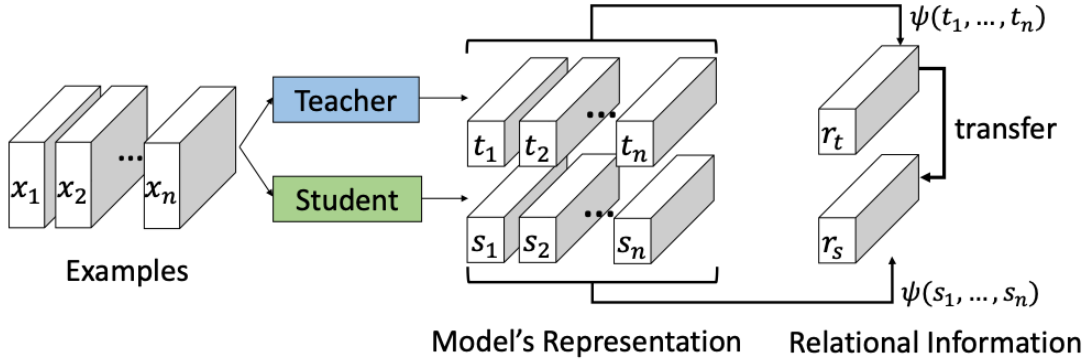
point to point -> group to group
自身属性建模 -> 关系建模

NeuHub
AI Links All

Relational KD



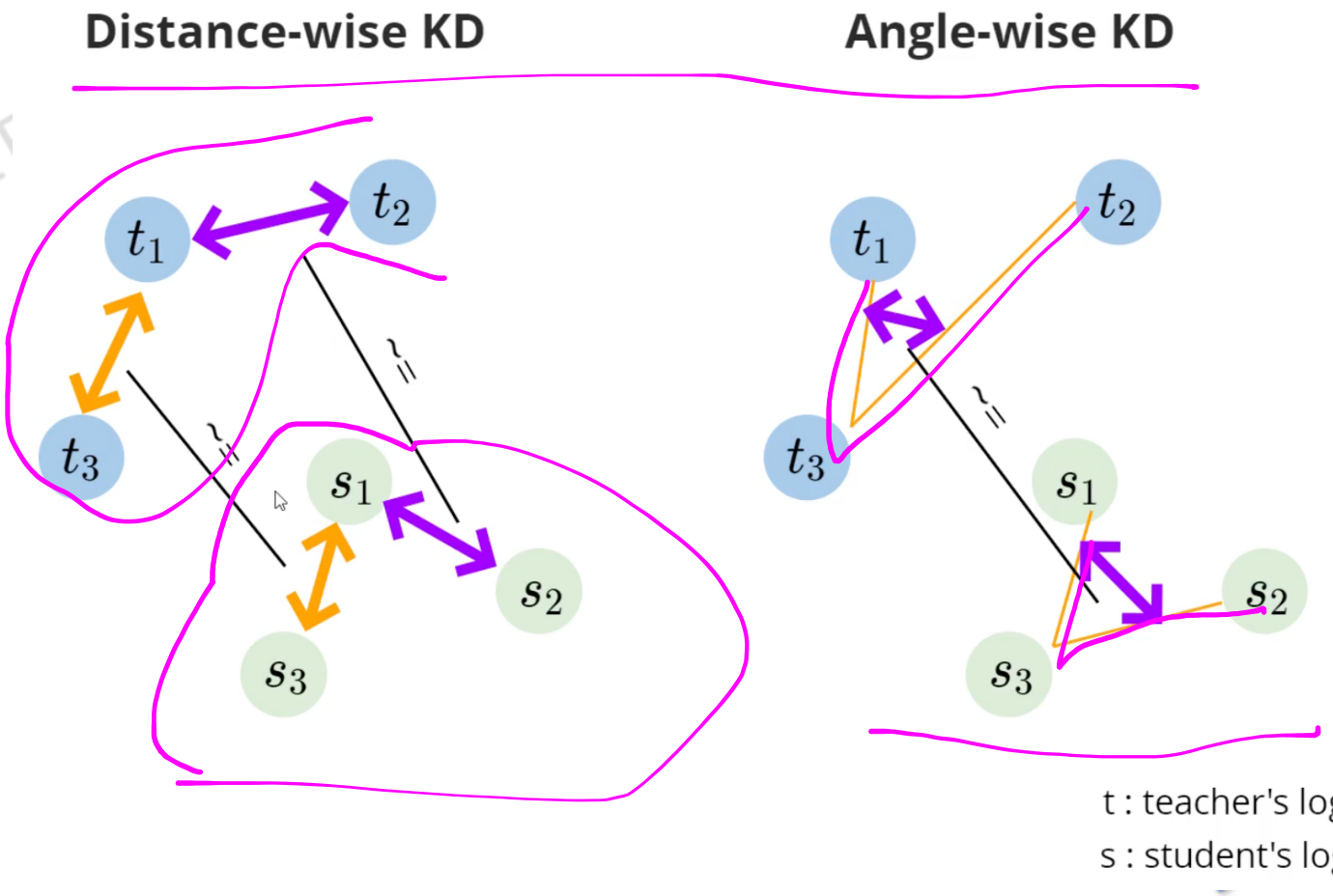
Individual Knowledge Distillation



Relational Knowledge Distillation

NeuHub
AI Links All

Relational KD



t : teacher's logits

s : student's logits

贪心科技 | 让每个人享受个性化教育服务

贪心科技 GREEDY TECHNOLOGY

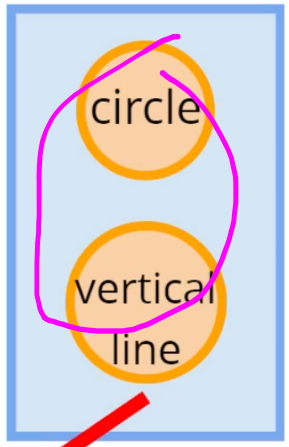
京东智联云 & 贪心科技 GREEDY TECHNOLOGY

Can we learn relations from feature ?
Similarity-Preserving Knowledge Distillation

NeuHub AI Links All

Similarity-Preserving Knowledge Distillation

Mnist Model



0
9
1

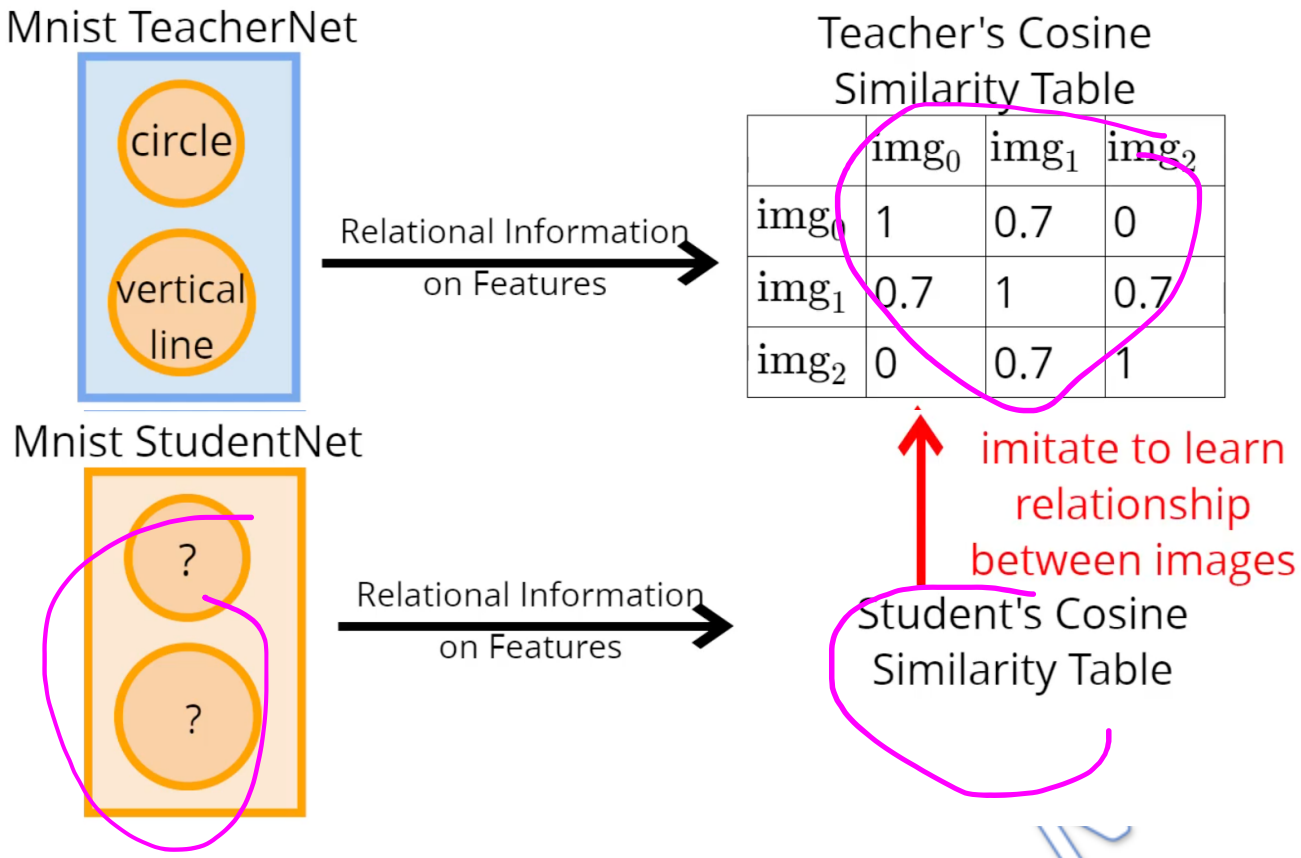
[1, 0]
[1, 1]
[0, 1]

Cosine Similarity

	img_0	img_1	img_2
img_0	1	0.7	0
img_1	0.7	1	0.7
img_2	0	0.7	1

Hub AI Links All

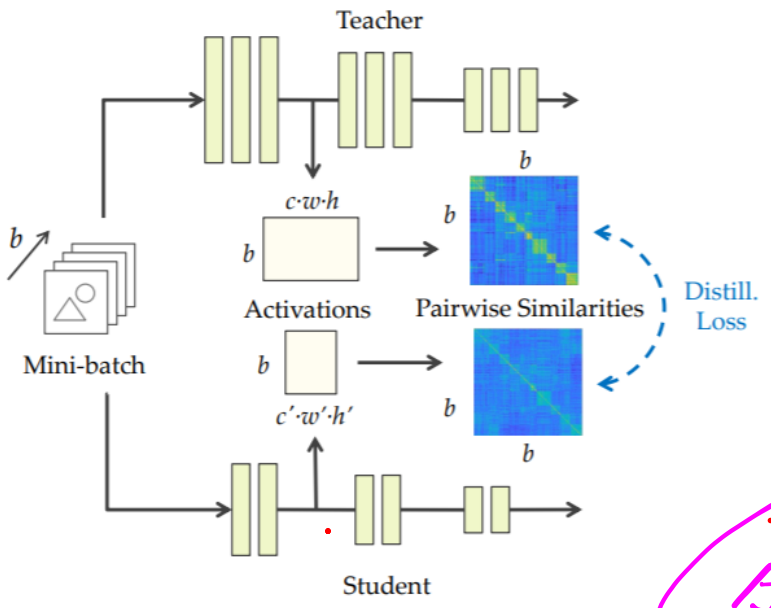
Similarity-Preserving Knowledge Distillation



	img_0	img_1	img_2
img_0	1	0.7	0
img_1	0.7	1	0.7
img_2	0	0.7	1

AI Links All

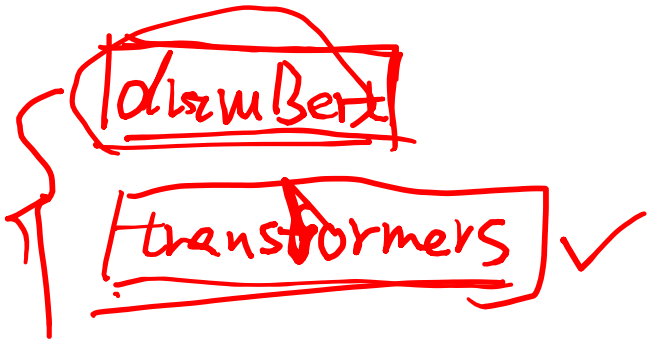
Similarity-Preserving Knowledge Distillation



NeuHub Links All

References

- https://zhuanlan.zhihu.com/p/102038521?utm_source=wechat_timeline
- https://arxiv.org/pdf/2006.05525.pdf
- https://zhuanlan.zhihu.com/p/90049906
- https://arxiv.org/abs/1412.6550
- https://arxiv.org/pdf/1904.05068v1.pdf
- https://arxiv.org/pdf/1805.05532.pdf



NeuHub
AI Links All